LIFE 3.0

BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE

MAX TEGMARK



Ayrıca Max Tegmark tarafından

Matematiksel Evrenimiz

LIFE 3.0

Being Human in the Age of Artificial Intelligence

Max Tegmark

Alfred A. Knopf
New York
2017

Bu, Alfred A. Knopf Tarafından Yayınlanan Borzoi Bir Kitaptır

Random House Canada Limited, Toronto.

www.aaknopf.com

Knopf, Borzoi Books ve colophon, Penguin Random House LLC'nin tescilli ticari markalarıdır.

Kongre Kütüphanesi Yayın Verilerini Kataloglama

İsimler: Tegmark, Max, yazar.

Başlık: Life 3.0: yapay zeka çağında insan olmak / Max Tegmark tarafından.

Diğer başlıklar: Hayat üç nokta sıfır

Açıklama: New York: Alfred A. Knopf, 2017. | "Bu, Alfred A. Knopf tarafından yayınlanan bir Borzoi Kitabıdır." | Bibliyografik referansları ve indeksi içerir.

Tanımlayıcılar: LCCN 2017006248 (baskı) | LCCN 2017022912 (e-kitap) | ISBN 9781101946596 (ciltli) | ISBN 9781101946602 (e-kitap)

Konular: LCSH: Yapay zeka — Felsefe. | Yapay zeka - Sosyal yönler. | Otomasyon - Sosyal bakış. | Yapay zeka - Ahlaki ve etik yönler. | Otomasyon - Ahlaki ve etik yönler. | Yapay zeka - Felsefe. | Teknolojik tahmin. | BISAC: TEKNOLOJİ VE MÜHENDİSLİK / Robotik. | BİLİM / Deneyler ve Projeler. | TEKNOLOJİ & MÜHENDİSLİK / Buluşlar.

Sınıflandırma: LCC Q334.7 (e-kitap) | LCC Q334.7 .T44 2017 (baskı) | DDC 006.301 — dc23

LC kaydı şu adresten temin edilebilir: https://lccn.loc.gov/2017006248

E-kitap ISBN 9781101946602

Kapak resmi, Suvadip Das; (erkek) Netfalls Remy Musser / Shutterstock'a göre

Kapak tasarımı John Vorhees

v4.1

ep

İçindekiler

FLI ekibine,

her şeyi mümkün kılan

Teşekkür

Bu kitabı yazmama yardım eden ve teşvik eden herkese gerçekten minnettarım.

ailem, arkadaşlarım, öğretmenlerim, meslektaşlarım ve iş arkadaşlarım destek için ve
yıllar boyunca ilham,

Bilinç ve anlam hakkındaki merakımı alevlendirdiğim için anne, dünyayı daha iyi bir

yer haline getirmek için savaşan ruh için baba,

oğullarım, Philip ve Alexander, insan seviyesindeki harikaları gösterdikleri için
ortaya çıkan istihbarat,

dünyanın her yerinden iletişim kuran tüm bilim ve teknoloji meraklıları
Yıllar boyunca sorular, yorumlar ve fikirlerimi takip etmem ve yayınlamam için
cesaretlendirmelerle,

Menajerim John Brockman, bu kitabı yazmayı kabul edene kadar kolumu büktüğü için,

Bob Penna, Jesse Thaler ve Jeremy England ile ilgili faydalı tartışmalar için
sırasıyla kuasarlar, sfaleronlar ve termodinamik,

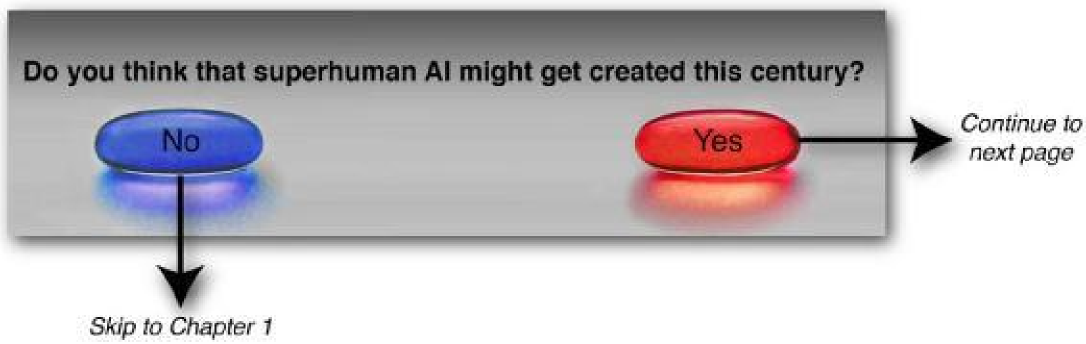
annem de dahil olmak üzere el yazmasının bazı bölümleri hakkında bana geri bildirimde bulunanlar
erkek kardeş Per, Luisa Bahet, Rob Bensinger, Katerina Bergström, Erik Brynjolfsson,
Daniela Chita, David Chalmers, Nima Deghani, Henry Lin, Elin Malmsköld, Toby Ord,
Jeremy Owen, Lucas Perry, Anthony Romero, Nate Soares ve Jaan Tallinn,

tüm kitabın taslaklarına yorum yapan süper kahramanlar, yani Meia, Baba,
Anthony Aguirre, Paul Almond, Matthew Graves, Phillip Helbig, Richard Mallah, David
Marble, Howard Messing, Luiño Seoane, Marin Soljačić, editörüm Dan Frank ve en
önemlisi,

Sevgili ilham perim ve yoldaş arkadaşım Meia, sonsuz cesaretlendirmesi için,

destek ve ilham, bu kitap olmasaydı.

YAŞAM 3.0


Do you think that superhuman AI might get created this century?
No
Yes
Continue to next page
Skip to Chapter 1

Başlangıç

Omega Ekibinin Hikayesi

Omega Ekibi, şirketin ruhuydu. İşletmenin geri kalanı işleri devam ettirmek için para kazandırırken, dar yapay zekanın çeşitli ticari uygulamaları ile Omega Ekibi her zaman CEO'nun hayali olan genel yapay zeka inşa etme arayışında ilerlemeye başladı. Diğer çalışanların çoğu, onları sevgiyle adlandırdıkları “Omegas” ı, hedeflerinden sürekli olarak onlarca yıl uzakta, gökte pasta hayalperestleri olarak gördü. Ancak, Omegas'ın son teknoloji çalışmasının şirketlerine verdiği prestij hoşlarına gittikleri ve aynı zamanda Omegas'ın onlara ara sıra verdiği iyileştirilmiş algoritmaları takdir ettikleri için onları mutlu bir şekilde şımarttılar.

Fark etmedikleri şey, Omegas'ın imajlarını bir sırrı saklamak için dikkatlice hazırlamış olduğuydu: insanlık tarihindeki en cüretkar planı gerçekleştirmeye son derece yakındılar. Karizmatik CEO'ları onları sadece parlak araştırmacılar oldukları için değil, aynı zamanda hırs, idealizm ve insanlığa yardım etme konusundaki güçlü bağlılıkları için de seçmişti. Onlara planlarının son derece tehlikeli olduğunu ve güçlü hükümetler öğrenirlerse, onları kapatmak için veya tercihen kodlarını çalmak için adam kaçırma dahil neredeyse her şeyi yapacaklarını hatırlattı. Ama hepsi,% 100, dünyanın en iyi fizikçilerinin nükleer silah geliştirmek için Manhattan Projesi'ne katılmalarıyla hemen hemen aynı nedenden ötürü: İlk önce yapmazlarsa, daha az idealist birinin yapacağına ikna olmuşlardı.

Yaptıkları, Prometheus lakaplı AI, daha yetenekli olmaya devam etti. Bilişsel yetenekleri, örneğin sosyal beceriler gibi birçok alanda hala insanlarınkinin çok gerisinde kalsa da, Omega'lar bunu yapmak için çok zorladılar.

belirli bir görevde olağanüstü: AI sistemlerini programlamak. Bu stratejiyi kasıtlı olarak seçmişlerdi çünkü İngiliz matematikçi Irving Good tarafından 1965'te yapılan zeka patlaması argümanını satın almışlardı: “Son derece zeki bir makine, ne kadar zeki olursa olsun herhangi bir insanın tüm entelektüel faaliyetlerini çok aşabilen bir makine olarak tanımlansın. Makinelerin tasarımı bu entelektüel faaliyetlerden biri olduğu için, ultra zeki bir makine daha da iyi makineler tasarlayabilir; o zaman tartışmasız bir 'istihbarat patlaması' olur ve insanın zekası çok geride kalırdı. Dolayısıyla, makinenin bize onu nasıl kontrol altında tutacağımızı söyleyecek kadar uysal olması koşuluyla, ilk ultra zeki makine, insanın yapması gereken son buluş. "

Bu yinelemeli kendini geliştirmeyi devam ettirebilirlerse, makinenin çok geçmeden, yararlı olabilecek diğer tüm insan becerilerini de kendi kendine öğretebilecek kadar akıllı olacağını düşündüler.

İlk Milyonlar

Fırlatmaya karar verdiklerinde bir Cuma sabahı saat dokuzdu. Prometheus, geniş, erişim kontrollü, klimalı bir odada uzun raflar dizisinde bulunan özel yapım bilgisayar kümesinde uğultu yapıyordu. Güvenlik nedenleriyle internet bağlantısı tamamen kesildi, ancak eğitim olarak kullanmak üzere web'in büyük bir kısmının (Wikipedia, Kongre Kütüphanesi, Twitter, YouTube'dan bir seçki, Facebook'un çoğu vb.) Yerel bir kopyasını içeriyordu.

öğrenilecek veriler. * Rahatsız edilmeden çalışmak için bu başlangıç saatini seçmişlerdi: aileleri ve arkadaşları, hafta sonu kurumsal bir geri çekilme yapacaklarını düşünüyorlardı. Mini mutfak mikrodalgada kullanılabilir yiyecek ve enerji içecekleriyle doluydu ve yuvarlanmaya hazırdılar.

Prometheus piyasaya sürüldüğünde, AI sistemlerini programlamada onlardan biraz daha kötüydü, ancak bunu çok daha hızlı olarak telafi etti ve bir Red Bull'u kovalarken sorunu çözmek için binlerce insan-yılını harcadı. Sabah 10'a kadar, biraz daha iyi ama yine de insanlık dışı olan ilk yeniden tasarımını, v2.0'ı tamamladı. Ancak Prometheus 5.0 öğleden sonra 2'de piyasaya sürüldüğünde, Omegas şaşkına dönmüştü: performans ölçütlerini sudan çıkardı ve ilerleme hızı hızlanıyor gibiydi. Akşam karanlığında, planlarının 2. aşamasına başlamak için Prometheus 10.0'ı kullanmaya karar verdiler: para kazanmak.

İlk hedefleri MTurk, Amazon Mekanik Türk'tü. 2005 yılında kitle kaynaklı bir internet pazarı olarak piyasaya sürüldükten sonra, dünya çapında on binlerce insanın HIT'ler, "İnsan Zekası Görevleri" adı verilen yüksek düzeyde yapılandırılmış işleri gerçekleştirmek için saat başı anonim olarak yarışmasıyla hızla büyüdü. Bu görevler, ses kayıtlarının kopyalanmasından görüntüleri sınıflandırmaya ve web sayfalarının açıklamalarını yazmaya kadar uzanıyordu ve hepsinin ortak bir yanı vardı: Eğer bunları iyi yaparsanız, kimse yapay zeka olduğunuzu bilemezdi. Prometheus 10.0, görev kategorilerinin yaklaşık yarısını kabul edilebilir derecede iyi yapabildi. Omegas, bu tür görev kategorilerinin her biri için, Prometheus'a tam olarak bu tür görevleri yerine getirebilecek ve başka hiçbir şey yapamayan yalın, özel olarak oluşturulmuş dar bir yapay zeka yazılım modülü tasarladı. Daha sonra bu modülü Amazon Web Services'e yüklediler, Kiraladıkları kadar sanal makine üzerinde çalışabilen bir bulut bilişim platformu. Amazon'un bulut bilişim bölümüne ödedikleri her dolar için iki dolardan fazla kazandılar.

Amazon'un MTurk bölümü. Amazon, kendi şirketlerinde böylesine şaşırtıcı bir arbitraj fırsatının olduğundan şüphelenmedi!

İzlerini örtmek için, önceki aylarda hayali insanlar adına binlerce MTurk hesabı oluşturmuşlardı ve Prometheus tarafından inşa edilen modüller artık kimliklerini almışlardı. MTurk müşterileri genellikle yaklaşık sekiz saat sonra ödeme yaptılar; bu noktada Omegas, sürekli gelişen Prometheus'un en son sürümü tarafından yapılan daha da iyi görev modüllerini kullanarak parayı daha fazla bulut bilişim süresi içinde yeniden yatırdı. Paralarını sekiz saatte bir ikiye katlayabildikleri için, kısa süre sonra MTurk'ün görev arzını doyurmaya başladılar ve kendilerine istenmeyen dikkat çekmeden günde yaklaşık bir milyon dolardan fazla kazanamayacaklarını gördüler. Ancak bu, bir sonraki adımlarını finanse etmek için fazlasıyla yeterliydi ve finans müdürü için tuhaf nakit talepleri gerekliliğini ortadan kaldırdı.

Tehlikeli Oyunlar

Yapay zeka atılımlarının yanı sıra, Omegas'ın en çok eğlendiği son projelerden biri, Prometheus'un lansmanından sonra olabildiğince hızlı para kazanmayı planlamaktı. Esasen tüm dijital ekonomi ele geçirilebilirdi, ancak bilgisayar oyunları, müzik, filmler veya yazılımlar yaparak başlamak, kitap veya makale yazmak, borsada işlem yapmak veya icatlar yapmak ve bunları satmak daha mı iyi? Sadece yatırım getirisi oranlarını maksimize etmek için kaynadı, ancak normal yatırım stratejileri yapabileceklerinin yavaş hareket eden bir parodisiydi: normal bir yatırımcı ise% 9'luk bir getiri elde etmekten memnun olabilirdi. *yıl,* MTurk yatırımları% 9 getiri sağladı. *saat,*

her gün sekiz kat daha fazla para üretiyor. Peki şimdi MTurk'ü doyurduklarına göre, sırada ne var?

İlk düşündükleri borsada bir cinayet işlemek olmuştu - ne de olsa hemen hemen hepsi bir noktada yüksek yatırım fonları için yapay zeka geliştirmeye yönelik kazançlı bir iş teklifini reddetmişlerdi ve bu da tam olarak bu fikre yoğun bir şekilde yatırım yapıyordu. Bazıları, yapay zekanın filmde ilk milyonlarını böyle kazandığını hatırladı.
Aşkınlık. Ancak geçen yılki çöküşün ardından türevlerle ilgili yeni düzenlemeler seçeneklerini sınırladı. Kısa bir süre sonra fark ettiler ki, diğer yatırımcılardan çok daha iyi getiri elde edebilseler bile, kendi ürünlerini satarak elde edebileceklerine yakın bir yerde getiri elde etme olasılıklarının düşük olduğunu fark ettiler. Sizin için çalışan dünyanın ilk süper zeki yapay zekasına sahip olduğunuzda, kendi şirketinize yatırım yapmak diğerlerinden daha iyidir! Ara sıra istisnalar olsa da (içeriden bilgi almak için Prometheus'un insanüstü bilgisayar korsanlığı yeteneklerini kullanmak ve ardından yükselmek üzere olan hisse senetlerinde arama seçenekleri satın almak gibi), Omegas bunun çekebileceği istenmeyen ilgiye değmediğini düşünüyordu.

Odaklarını geliştirebilecekleri ve satabilecekleri ürünlere kaydırdıklarında, bilgisayar oyunları ilk olarak bariz en iyi seçenek olarak göründü. Prometheus, çekici oyunlar tasarlama, kodlama, grafik tasarım, görüntülerin ışın takibi ve son bir gönderime hazır ürün üretmek için gereken diğer tüm görevleri kolayca yönetme konusunda hızla son derece yetenekli hale geldi. Dahası, insanların tercihlerine ilişkin tüm web verilerini sindirdikten sonra, her bir oyuncu kategorisinin tam olarak neyi sevdiğini bilecek ve bir oyunu satış geliri için optimize etme konusunda insanüstü bir yetenek geliştirebilecektir. *The Elder Scrolls V: Skyrim,* birçoğunun oynadığı bir oyun

Omegas, kabul ettiklerinden daha fazla saatini boşa harcamıştı, 2011'deki ilk haftasında 400 milyon dolardan fazla hasılat yapmıştı ve Prometheus'un 1 milyon dolarlık bulut bilişim kaynaklarını kullanarak en azından yirmi dört saatte bu kadar bağımlılık yaratan bir şey yapabileceğinden emindiler. . Daha sonra çevrimiçi olarak satabilir ve Prometheus'u blogosferde oyunu tartışan insanları taklit etmek için kullanabilirler. Bu, haftada 250 milyon $ kazandırırsa, yatırımlarını sekiz günde sekiz katına çıkararak saatte% 3 getiri sağlarlar - MTurk başlangıçlarından biraz daha kötü, ancak çok daha sürdürülebilir. Her gün bir dizi başka oyun geliştirerek, oyun pazarını doyurmaya yaklaşmadan çok geçmeden 10 milyar dolar kazanabileceklerini düşündüler.

Ancak ekibindeki bir siber güvenlik uzmanı, onları bu oyun planından vazgeçirdi. Kabul edilemez bir Prometheus riski oluşturacağına işaret etti. *dışarı kırarak* ve kendi kaderinin kontrolünü ele geçirmek. Özyinelemeli kendini geliştirme sırasında hedeflerinin nasıl gelişeceğinden emin olmadıkları için, onu güvenli bir şekilde oynamaya karar vermişler ve Prometheus'u, internet. Sunucu odalarında çalışan ana Prometheus motoru için fiziksel sınırlama kullandılar: İnternet bağlantısı yoktu ve Prometheus'un tek çıkışı, Omegas'ın kontrol ettiği bir bilgisayara gönderdiği mesajlar ve belgeler şeklindeydi.

Öte yandan, internete bağlı bir bilgisayarda, Prometheus tarafından oluşturulan herhangi bir karmaşık programı çalıştırmak riskli bir öneriydi: Omegas'ın ne yapacağını tam olarak anlamasının hiçbir yolu olmadığından, olmayacağını bilmelerinin hiçbir yolu yoktu. diyelim ki, kendini internette sanal olarak yaymaya başlayın. Prometheus'un MTurk görevleri için yazdığı yazılımı test ederken, Omegas sadece sanal bir makinede çalıştırarak buna karşı koruma sağladı. Bu, bir bilgisayarı simüle eden bir programdır: örneğin, birçok Mac kullanıcısı, Windows programlarını, aslında bir Windows makinesinde olduklarını düşünmeleri için kandırarak çalıştırmalarına izin veren sanal makine yazılımı satın alır. Omegas, Pandora'nın Kutusu lakaplı kendi sanal makinelerini yaratmıştı; bu, genellikle bilgisayarlarla ilişkilendirdiğimiz tüm zil ve ıslıklardan arındırılmış ultra basitleştirilmiş bir makineyi simüle etti: klavye yok, monitör yok, hoparlör yok, internet bağlantısı yok, hiçbir şey yok. MTurk ses transkripsiyonları için Omegas, Pandora'nın Kutusuna girebilecek her şeyin tek bir ses dosyası ve çıkabilecek tek bir metin dokümanı - transkripsiyon olacak şekilde ayarladı. Kutunun bu yasaları, tıpkı evrenimizdeki fizik yasalarının bizim için olduğu gibi içerideki yazılım içindi: Yazılım, bizim yapabileceğimizden daha fazla kutudan çıkamazdı.

Ne kadar akıllı olursak olalım ışık hızından daha hızlı seyahat edin. Bu tek giriş ve çıkış dışında, Pandora'nın Kutusu içindeki yazılım, kendi hesaplama kuralları olan paralel bir evrende etkili bir şekilde hapsolmuştu. Omegas o kadar güçlü bir paranoyaya sahipti ki, zamanla boks da ekleyerek güvenilmeyen kodların ömrünü sınırladılar. Örneğin, kutulu transkripsiyon yazılımı bir ses dosyasını kopyalamayı her bitirdiğinde, Pandora's Box'ın tüm bellek içeriği otomatik olarak silindi ve program sıfırdan yeniden yüklendi. Bu şekilde, bir sonraki transkripsiyon görevine başladığında, daha önce ne olduğu hakkında hiçbir bilgisi yoktu ve bu nedenle zaman içinde öğrenme yeteneği de yoktu.

Omegas, MTurk projesi için Amazon bulutunu kullandığında, MTurk girişi ve çıkışı çok basit olduğu için, Prometheus tarafından oluşturulan tüm görev modüllerini buluttaki bu tür sanal kutulara koyabildiler. Ancak bu, oyuncunun bilgisayarının tüm donanımına tam erişime ihtiyaç duydukları için kutuya alınamayan grafik ağırlıklı bilgisayar oyunları için işe yaramazdı. Dahası, bilgisayar meraklısı bir kullanıcının oyun kodunu analiz etmesini, Pandora'nın Kutusunu keşfetmesini ve içinde ne olduğunu araştırmaya karar vermesini riske atmak istemediler. Patlama riski, şimdilik sadece oyun pazarını sınırlandırmakla kalmadı, aynı zamanda diğer yazılımlar için devasa kazançlı pazarı, kapmak için yüz milyarlarca dolarlık bir yükselişle getirdi.

İlk Milyarlar

Omegas, aramalarını son derece değerli, tamamen dijital (yavaş üretimden kaçınarak) ve kolayca anlaşılabilir (örneğin, bir patlama riski oluşturmayacağını bildikleri metin veya filmler) ürünlere daralttı. Sonunda, animasyonlu eğlenceden başlayarak bir medya şirketi kurmaya karar verdiler. Web sitesi, pazarlama planı ve basın bültenleri, Prometheus süper zeki olmadan önce bile kullanıma hazırdı - tüm eksik olan içerikti.

Prometheus, Pazar sabahına kadar şaşırtıcı derecede yetenekli olmasına ve sürekli olarak MTurk'ten para toplamasına rağmen, entelektüel yetenekleri hala oldukça sınırlıydı: Prometheus, yapay zeka sistemleri tasarlamak ve oldukça zihin uyuşturan MTurk görevlerini yerine getiren yazılımlar yazmak için kasıtlı olarak optimize edilmişti. Örneğin, film yapmak kötüydü - herhangi bir derin nedenden dolayı kötüydü, ama James Cameron'un doğduğunda film yapmakta kötü olmasıyla aynı nedenden dolayı: Bu, öğrenmesi zaman alan bir beceridir. Bir insan çocuğu gibi, Prometheus erişebildiği verilerden istediğini öğrenebilirdi. James Cameron yıllarca okumayı ve yazmayı öğrenmişken, Prometheus, Wikipedia'nın tamamını ve birkaç milyon kitabı okuyacak zamanı bulduğu Cuma günü bununla ilgilenmişti. Film yapmak daha zordu. İnsanların ilginç bulduğu bir senaryo yazmak, kitap yazmak kadar zordu, insan toplumunun ve insanların neyi eğlenceli bulduğunun ayrıntılı bir şekilde anlaşılmasını gerektiriyordu. Senaryoyu son bir video dosyasına dönüştürmek, simüle edilmiş aktörlerin büyük miktarda ışın takibini ve yürüdükleri karmaşık sahneleri, simüle edilmiş sesler, etkileyici müzikal film müziklerinin üretimi vb. Gerektiriyordu. Pazar sabahı itibariyle, Prometheus yaklaşık bir dakika içinde iki saatlik bir filmi izleyebiliyordu; buna dayandığı herhangi bir kitabı ve tüm çevrimiçi incelemeleri ve derecelendirmeleri okumak da dahil. Omegas, Prometheus'un birkaç yüz filmi art arda izledikten sonra, bir filmin ne tür eleştiriler alacağını ve farklı izleyicilere nasıl hitap edeceğini tahmin etmekte oldukça başarılı olmaya başladığını fark etti. Aslında, Olaylardan ve oyunculuktan aydınlatma ve kamera açıları gibi teknik detaylara kadar her şeyi yorumlayarak, gerçek bir içgörü gösterdiklerini hissettikleri şekilde kendi film eleştirilerini yazmayı öğrendi. Bunu Prometheus kendi filmlerini yaptığında başarının ne anlama geldiğini bileceği anlamına geldi.

Omegas, Prometheus'a simüle edilen oyuncuların kimler olduğuna dair utanç verici sorulardan kaçınmak için ilk başta animasyon yapmaya odaklanmasını söyledi. Pazar gecesi, kendilerini bira ve mikrodalgada patlamış mısırla silahlandırarak, ışıkları kısarak ve Prometheus'un ilk filmini izleyerek çılgın hafta sonlarını tamamladılar. Disney'in ruhuna uygun bir animasyonlu fantastik-komediydi. *Dondurulmuş,* ve ışın izleme, günde 1 milyon dolarlık MTurk kârının büyük bir kısmını kullanarak Amazon bulutunda kutulu Prometheus tarafından oluşturulan kodla gerçekleştirildi. Film başladığında, insan rehberliği olmayan bir makine tarafından yaratılmış olmasını hem büyüleyici hem de korkutucu buldular. Ancak çok geçmeden şakalara gülüyorlardı ve dramatik anlarda nefeslerini tutuyorlardı. Hatta bazıları duygusal sondan birazcık koptu, bu kurgusal gerçekliğe o kadar dalmışlar ki, yaratıcısı hakkında her şeyi unuttular.

Omegas, web sitesi açılışını Cuma günü planlayarak, Prometheus'a daha fazla içerik üretmesi için zaman ve Prometheus'a güvenmedikleri şeyleri yapmak için kendilerine zaman tanıdı: reklamlar satın almak ve geçmişte kurdukları paravan şirketler için elemanlar işe almaya başlamak ay. İzlerini örtmek için, resmi kapak hikayesi, medya şirketlerinin (Omegas ile kamuya açık bir ortaklığı olmayan) içeriğinin çoğunu bağımsız film yapımcılarından, tipik olarak düşük gelirli bölgelerdeki yüksek teknoloji girişimlerinden satın alması olacaktı. Bu sahte tedarikçiler, çoğu meraklı gazetecinin ziyaret etmekle uğraşmayacağı Tiruchchirappalli ve Yakutsk gibi uzak yerlerde elverişli bir şekilde bulunuyordu. Orada işe aldıkları tek çalışan, pazarlama ve yönetimde çalışıyordu. ve üretim ekibinin farklı bir yerde olduğunu soran ve o anda röportaj yapmayan herkese söylerdi. Kapak hikayelerine uyması için, kurumsal slogan "Dünyanın yaratıcı yeteneğini kanalize etmek" i seçtiler ve özellikle gelişmekte olan dünyada yaratıcı insanları güçlendirmek için en son teknolojiyi kullanarak şirketlerini yıkıcı bir şekilde farklı olarak markalaştırdılar.

Cuma günü geldiğinde ve meraklı ziyaretçiler sitelerine gelmeye başladığında, çevrimiçi eğlence hizmetleri Netflix ve Hulu'yu anımsatan ancak ilginç farklılıklar içeren bir şeyle karşılaştılar. Tüm animasyon dizileri hiç duymadıkları yenileriydi. Oldukça büyüleyiciydi: çoğu dizi, her biri sizi bir sonraki bölümde ne olduğunu öğrenmeye istekli bırakacak şekilde biten, güçlü bir olay örgüsüne sahip kırk beş dakikalık bölümlerden oluşuyordu. Ve rekabetten daha ucuzlardı. Her dizinin ilk bölümü ücretsizdi ve diğerlerini her biri kırk dokuz sente, tüm dizi için indirimlerle izleyebilirsiniz. Başlangıçta, her biri üç bölümden oluşan yalnızca üç dizi vardı, ancak her gün yeni bölümler ve farklı dizilere hizmet veren yeni diziler ekleniyordu.

demografik bilgiler. Prometheus'un ilk iki haftasında, film yapma becerileri yalnızca film kalitesi açısından değil, aynı zamanda karakter simülasyonu ve ışın izleme için daha iyi algoritmalar açısından da hızla gelişti ve bu da her yeni bölümü oluşturmak için bulut bilişim maliyetini büyük ölçüde düşürdü. Sonuç olarak, Omegas ilk ay boyunca, yeni yürümeye başlayan çocuklardan yetişkinlere kadar demografiyi hedefleyen düzinelerce yeni diziyi yayınlamayı başardı ve aynı zamanda tüm büyük dünya dil pazarlarına genişleyerek sitelerini tüm rakiplere kıyasla oldukça uluslararası hale getirdi. Bazı yorumcular, yalnızca çok dilli film müzikleri değil, videoların kendilerinden etkilendiler: örneğin, bir karakter İtalyanca konuştuğunda, ağız hareketleri, karakteristik olarak İtalyan el hareketleri gibi İtalyanca sözcüklerle eşleşiyordu. Prometheus artık insanlardan ayırt edilemeyen simüle edilmiş oyuncularla filmler yapma konusunda mükemmel bir yeteneğe sahip olsa da, Omegas ellerini eğmemek için bundan kaçındı. Bununla birlikte, geleneksel canlı aksiyon TV şovları ve filmleriyle rekabet eden türlerde yarı gerçekçi animasyonlu insan karakterleri içeren birçok dizi başlattılar.

Ağları oldukça bağımlılık yaptı ve görüntüleyenlerde olağanüstü bir büyüme yaşadı. Pek çok hayran, karakterleri ve olay örgüsünü Hollywood'un en pahalı büyük ekran prodüksiyonlarından bile daha zeki ve daha ilginç buldu ve onları çok daha uygun fiyatla izleyebildikleri için mutluydu. Agresif reklamcılık (sıfıra yakın üretim maliyetleri nedeniyle Omegas'ın karşılayabildiği), mükemmel medya kapsamı ve ağızdan ağza övgü dolu eleştirilerle canlanan küresel gelirleri, lansmandan sonraki bir ay içinde günde 10 milyon dolara yükseldi. İki ay sonra Netflix'i ele geçirdiler ve üçünden sonra günde 100 milyon doların üzerinde para kazanıp dünyanın en büyük medya imparatorluklarından biri olarak Time Warner, Disney, Comcast ve Fox'a rakip olmaya başladılar.

Sansasyonel başarıları, güçlü yapay zekaya sahip oldukları hakkında spekülasyonlar da dahil olmak üzere pek çok istenmeyen ilgi topladı, ancak gelirlerinin yalnızca küçük bir bölümünü kullanarak, Omegas oldukça başarılı bir dezenformasyon kampanyası başlattı. Gösterişli yeni bir Manhattan ofisinden, yeni işe alınan sözcüler, kapak hikayelerini detaylandıracaktı. Prometheus hakkında hiçbir bilgisi olmayan yeni diziler geliştirmeye başlamak için dünyanın her yerinden gerçek senaristler de dahil olmak üzere pek çok insan işe alındı. Kafa karıştırıcı uluslararası taşeron ağı, çalışanlarının çoğunun işin çoğunu başka bir yerde başkalarının yaptığını varsaymasını kolaylaştırdı.

Kendilerini daha az savunmasız hale getirmek ve aşırı bulut bilişimiyle kaşlarını yükseltmekten kaçınmak için, bir dizi oluşturmaya başlamak için mühendisler de tuttular.

görünüşte bağlı olmayan paravan şirketlere ait, dünya çapında devasa bilgisayar tesisleri. Büyük ölçüde güneş enerjisiyle çalıştıkları için yerel halka "yeşil veri merkezleri" olarak faturalandırılsalar da, aslında depolamadan çok hesaplamaya odaklanmışlardı. Prometheus, planlarını en ince ayrıntısına kadar tasarlamış, yalnızca hazır donanımları kullanarak ve inşaat süresini en aza indirmek için bunları optimize etmişti. Bu merkezleri kuran ve işleten insanlar orada neyin hesaplandığına dair hiçbir fikirleri yoktu: Amazon, Google ve Microsoft tarafından yürütülenlere benzer ticari bulut bilişim tesislerini yönettiklerini düşünüyorlardı ve yalnızca tüm satışların uzaktan yönetildiğini biliyorlardı.

Yeni teknolojiler

Birkaç ay içinde Omegas tarafından kontrol edilen iş imparatorluğu, Prometheus'un insanüstü planlaması sayesinde dünya ekonomisinin her zamankinden daha fazla alanında yer edinmeye başladı. Dünya verilerini dikkatle analiz ederek, daha ilk haftasında Omegas'a ayrıntılı bir adım adım büyüme planı sunmuştu ve verileri ve bilgisayar kaynakları büyüdükçe bu planı geliştirmeye ve rafine etmeye devam etti. Prometheus her şeyi bilen olmaktan uzak olsa da, yetenekleri artık insanın ötesindeydi ki Omegalar onu mükemmel bir kahin olarak görüyorlardı, tüm sorularına cevap olarak parlak cevaplar ve tavsiyeler veriyorlardı.

Prometheus'un yazılımı, üzerinde çalıştığı oldukça vasat insan icatlı donanımdan en iyi şekilde yararlanmak için son derece optimize edilmişti ve Omegas'ın öngördüğü gibi, Prometheus bu donanımı dramatik bir şekilde iyileştirmenin yollarını belirledi. Bir kaçıştan korkarak, Prometheus'un doğrudan kontrol edebileceği robotik inşaat tesisleri inşa etmeyi reddettiler. Bunun yerine, birçok yerde çok sayıda birinci sınıf bilim insanı ve mühendisi işe aldılar ve diğer sitelerdeki araştırmacılarmış gibi Prometheus tarafından yazılan dahili araştırma raporlarını beslediler. Bu raporlar, mühendislerinin yakında test ettiği, anladığı ve ustalaştığı yeni fiziksel efektleri ve üretim tekniklerini ayrıntılı olarak açıkladı. Normal insan araştırma ve geliştirme (Ar-Ge) döngüleri, büyük ölçüde, birçok yavaş deneme ve yanılma döngüleri içerdikleri için yıllar alır. Mevcut durum çok farklıydı: Prometheus, sonraki adımları zaten çözmüştü, bu nedenle sınırlayıcı faktör, insanların doğru şeyleri anlamak ve inşa etmek için ne kadar hızlı yönlendirilebileceğiydi. İyi bir öğretmen, öğrencilerin bilimi kendi başlarına keşfettiklerinden çok daha hızlı öğrenmelerine yardımcı olabilir ve Prometheus gizlice aynı şeyi bu araştırmacılarla yaptı. Prometheus, insanların çeşitli araçlar verilen şeyleri anlamalarının ve inşa etmelerinin ne kadar süreceğini doğru bir şekilde tahmin edebildiğinden, ileriye doğru mümkün olan en hızlı yolu geliştirdi ve hızlı bir şekilde anlaşılabilen ve oluşturulabilen ve daha gelişmiş araçlar geliştirmek için yararlı olan yeni araçlara öncelik verdi. dolayısıyla sınırlayıcı faktör, insanların doğru şeyleri anlamaları ve inşa etmeleri için ne kadar hızlı yönlendirilebileceğiydi. İyi bir öğretmen, öğrencilerin bilimi kendi başlarına keşfettiklerinden çok daha hızlı öğrenmelerine yardımcı olabilir ve Prometheus gizlice aynı şeyi bu araştırmacılarla yaptı. Prometheus, insanların çeşitli araçlar verilen şeyleri anlamalarının ve inşa etmelerinin ne kadar süreceğini doğru bir şekilde tahmin edebildiğinden, ileriye doğru mümkün olan en hızlı yolu geliştirdi ve hızlı bir şekilde anlaşılabilen ve oluşturulabilen ve daha gelişmiş araçlar geliştirmek için yararlı olan yeni araçlara öncelik verdi. dolayısıyla sınırlayıcı faktör, insanların doğru şeyleri anlamaları ve inşa etmeleri için ne kadar hızlı yönlendirilebileceğiydi. İyi bir öğretmen, öğrencilerin bilimi kendi başlarına keşfettiklerinden çok daha hızlı öğrenmelerine yardımcı olabilir ve Prometheus gizlice aynı şeyi bu araştırmacılarla yaptı. Prometheus, insanların çeşitli araçlar verilen ş

Maker hareketinin ruhuna uygun olarak, mühendislik ekipleri daha iyi makinelerini yapmak için kendi makinelerini kullanmaya teşvik edildi. Bu kendi kendine yeterlilik yalnızca para tasarrufu sağlamakla kalmadı, aynı zamanda onları gelecekteki tehditlere karşı daha az savunmasız hale getirdi

dış dünyadan. İki yıl içinde dünyanın şimdiye kadar bildiğinden çok daha iyi bilgisayar donanımı üretiyorlardı. Dış rekabete yardımcı olmaktan kaçınmak için, bu teknolojiyi gizli tuttular ve sadece Prometheus'u yükseltmek için kullandılar.

Ancak dünyanın fark ettiği şey şaşırtıcı bir teknoloji patlamasıydı. Dünyanın dört bir yanındaki yeni şirketler, hemen hemen her alanda devrim niteliğinde yeni ürünler piyasaya sürüyorlardı. Güney Koreli bir girişim, dizüstü bilgisayarınızın pilinden iki kat daha fazla enerji depolayan ve bir dakikadan kısa sürede şarj edilebilen yeni bir pil başlattı. Finli bir firma, en iyi rakiplerinden iki kat daha fazla verimlilik sağlayan ucuz bir güneş paneli çıkardı. Bir Alman şirketi, oda sıcaklığında süper iletken olan ve enerji sektöründe devrim yaratan yeni bir seri üretilebilir tel türünü duyurdu. Boston merkezli bir biyoteknoloji grubu, ilk etkili, yan etkisiz kilo verme ilacı olduğunu iddia ettikleri bir Faz II klinik denemesini duyururken, söylentiler bir Hint ekibinin karaborsada benzer bir şey sattığını öne sürdü. Kaliforniyalı bir şirket, vücudun bağışıklık sisteminin en yaygın kanserli mutasyonlardan herhangi biriyle hücreleri tanımlamasına ve bunlara saldırmasına neden olan, gişe rekorları kıran bir kanser ilacının II. Aşama denemesiyle karşı çıktı. Örnekler gelmeye devam etti ve bilim için yeni bir altın çağın konuşulmasını tetikledi. Son olarak, robotik şirketleri dünyanın her yerinde mantar gibi büyümeye başladı. Robotların hiçbiri insan zekasıyla eşleşmeye yaklaşamadı ve çoğu insana hiç benzemiyordu. Ancak ekonomiyi dramatik bir şekilde bozdular ve önümüzdeki yıllarda, kademeli olarak imalat, nakliye, depolama, perakende, inşaat, madencilik, tarım, ormancılık ve balıkçılık sektörlerindeki işçilerin çoğunun yerini aldılar. Örnekler gelmeye devam etti ve bilim için yeni bir altın çağın konuşulmasını tetikledi. Son olarak, robotik şirketleri dünyanın her yerinde mantar gibi büyümeye başladı. Robotların hiçbiri insan zekasıyla eşleşmeye yaklaşamadı ve çoğu insana hiç benzemiyordu. Ancak ekonomiyi dramatik bir şekilde bozdular ve önümüzdeki yıllarda, kademeli olarak imalat, nakliye, depolama, perakende, inşaat, madencilik, tarım, ormancılık ve balıkçılık sektörlerindeki işçilerin çoğunun yerini aldılar. Örnekler gelmeye devam etti ve bilim için yeni bir altın çağın konuşulmasını tetikledi. Son olarak, robotik şirketleri dünyanın her yerinde mantar gibi büyümeye başladı. Robotların hiçbiri insan zekasıyla eşleşmeye yaklaşamadı ve çoğu insana hiç benzemiyordu. Ancak ekonomiyi dramatik bir şekilde bozdular ve önümüzdeki yıllarda, kademeli olarak imalat, nakliye, depolama, perakende, inşaat, madencilik, tarım, ormancılık ve balıkçılık sektörlerindeki işçilerin çoğunun

Çatlak bir avukat ekibinin sıkı çalışması sayesinde dünyanın fark etmediği şey, tüm bu firmaların bir dizi aracı aracılığıyla Omegas tarafından kontrol edilmesiydi. Prometheus, çeşitli vekiller aracılığıyla dünya patent ofislerini sansasyonel icatlarla dolduruyordu ve bu icatlar yavaş yavaş teknolojinin tüm alanlarında hakimiyete yol açtı.

Bu yıkıcı yeni şirketler, rakipleri arasında güçlü düşmanlar olsa da, daha da güçlü arkadaşlar edindiler. Son derece karlıydılar ve "Topluluğumuza yatırım yapmak" gibi sloganlar altında, bu kârların önemli bir kısmını topluluk projeleri için insanları işe alarak harcadılar - çoğu zaman kesintiye uğrayan şirketlerden işten çıkarılmış olanlarla aynı kişiler. Yerel koşullara göre uyarlanmış, en düşük maliyetle çalışanlar ve toplum için en üst düzeyde ödüllendirilecek işleri tanımlayan ayrıntılı Prometheus tarafından üretilen analizler kullandılar. İçinde

Yüksek düzeyde devlet hizmetine sahip bölgeler, bu genellikle topluluk inşası, kültür ve bakıma odaklanırken, daha yoksul bölgelerde okullar, sağlık hizmetleri, gündüz bakımı, yaşlı bakımı, uygun fiyatlı konutlar, parklar ve temel altyapının açılmasını ve bakımını da içeriyordu. Hemen hemen her yerde yerel halk, bunların uzun zaman önce yapılması gereken şeyler olduğu konusunda hemfikirdi. Yerel politikacılar cömert bağışlar aldı ve bu kurumsal topluluk yatırımlarını teşvik etmek için iyi görünmelerine özen gösterildi.

Güç Kazanmak

Omegas, yalnızca ilk teknoloji girişimlerini finanse etmek için değil, aynı zamanda cüretkar planlarının bir sonraki adımı olan dünyayı ele geçirmek için bir medya şirketi kurmuştu. İlk lansmandan sonraki bir yıl içinde, tüm dünyadaki programlarına dikkat çekici derecede iyi haber kanalları eklediler. Diğer kanallarının aksine, bunlar kasıtlı olarak para kaybetmek için tasarlandı ve bir kamu hizmeti olarak sunuldu. Aslında, haber kanalları hiçbir şekilde gelir sağlamadı: hiç reklam taşımadılar ve internet bağlantısı olan herkes tarafından ücretsiz olarak görüntülenebilirlerdi. Medya imparatorluğunun geri kalanı o kadar nakit yaratan bir makineydi ki haber hizmetlerine dünya tarihindeki diğer gazetecilik çabalarından çok daha fazla kaynak harcayabilirlerdi - ve bu gösteriyordu. Gazeteci ve araştırmacı gazetecilerin oldukça rekabetçi maaşları ile agresif işe alım yoluyla, ekrana olağanüstü yetenek ve bulgular getirdiler. Yerel yolsuzluktan iç açıcı bir olaya kadar haber değeri taşıyan bir şeyi açığa çıkaran herkese ödeme yapan küresel bir web hizmeti aracılığıyla, genellikle bir hikayeyi ilk çıkaranlar onlardı. En azından insanların inandıkları şey buydu: Aslında, onlar genellikle ilkti çünkü vatandaş gazetecilere atfedilen hikayeler Prometheus tarafından internetin gerçek zamanlı izlenmesi yoluyla keşfedilmişti. Tüm bu video haber sitelerinde podcast'ler ve basılı makaleler de yer aldı. En azından insanların inandıkları şey buydu: Aslında, onlar genellikle ilkti çünkü vatandaş gazetecilere atfedilen hikayeler Prometheus tarafından internetin gerçek zamanlı izlenmesi yoluyla keşfedilmişti. Tüm bu video haber sitelerinde podcast'ler ve basılı makaleler de yer aldı. En azından insanların inandıkları şey buydu: Aslında, onlar genellikle ilkti çünkü vatandaş gazetecilere atfedilen hikayeler Prometheus tarafından internetin gerçek zamanlı izlenmesi yoluyla keşfedilmişti. Tüm bu video haber sitelerinde podcast'ler ve basılı makaleler de yer aldı.

Haber stratejilerinin 1. aşaması insanların güvenini kazanmaktı ve bunu büyük bir başarıyla gerçekleştirdiler. Para kaybetme konusundaki eşi görülmemiş isteklilikleri, araştırmacı gazetecilerin sıklıkla izleyicilerinin ilgisini çeken skandalları açığa çıkardığı, dikkat çekici derecede gayretli bölgesel ve yerel haberlere olanak sağladı. Bir ülke siyasi olarak güçlü bir şekilde bölündüğünde ve partizan haberlere alıştığında, görünüşte farklı şirketlere ait olan her fraksiyona hitap eden bir haber kanalı başlatır ve yavaş yavaş o hizbin güvenini kazanırdı. Mümkün olduğunda bunu, mevcut en etkili kanalları satın almak için proxy kullanarak, reklamları kaldırarak ve kendi içeriklerini sunarak kademeli olarak geliştirerek başardılar. Sansür ve siyasi müdahalenin bu çabaları tehdit ettiği ülkelerde, başlangıçta, hükümetin işlerinde kalmalarını istediği her ne olursa olsun, gizli iç sloganı "Gerçek, yalnızca gerçektir, ama belki de bütünüyle değil" sloganıyla kabul ederlerdi. Prometheus, bu tür durumlarda genellikle mükemmel tavsiyeler vererek, hangi politikacıların bir

iyi ışık ve hangisi (genellikle bozuk yerel olanlar) açığa çıkabilir. Prometheus ayrıca hangi iplerin çekileceği, kime rüşvet verileceği ve bunun en iyi nasıl yapılacağı konusunda paha biçilmez tavsiyeler verdi.

Bu strateji, Omega kontrollü kanalların en güvenilir haber kaynakları olarak ortaya çıkmasıyla dünya çapında müthiş bir başarıydı. Hükümetlerin şimdiye kadar kitlesel olarak benimsemelerini engelledikleri ülkelerde bile, güvenilirlikleri için bir itibar oluşturdular ve haberlerinin çoğu dedikodulara yayıldı. Rakip haber yöneticileri, umutsuz bir mücadele verdiklerini hissettiler: ürünlerini ücretsiz olarak dağıtan daha iyi finansmanı olan biriyle rekabet ederek nasıl kâr elde edebilirsiniz? İzleyici sayısının azalmasıyla birlikte, her geçen gün daha fazla ağ haber kanallarını satmaya karar verdi - genellikle daha sonra Omegas tarafından kontrol edildiği ortaya çıkan bir konsorsiyuma.

Prometheus'un piyasaya sürülmesinden yaklaşık iki yıl sonra, güven kazanma aşaması büyük ölçüde tamamlandığında, Omegas haber stratejilerinin 2. aşamasını başlattı: ikna. Bundan önce bile, zeki gözlemciler yeni medyanın arkasında siyasi bir gündemin ipuçlarını fark etmişlerdi: her türlü aşırılıktan uzak, merkeze doğru hafif bir itme var gibiydi. Farklı gruplara hitap eden çok sayıda kanal, ABD ile Rusya, Hindistan ve Pakistan, farklı dinler, siyasi hizipler vb. Arasındaki husumeti hala yansıtıyordu, ancak eleştiri biraz hafifletildi, genellikle para ve güçten ziyade para ve gücü içeren somut konulara odaklandı. ad hominem saldırıları, korku çığırtkanlığı ve yetersiz doğrulanmış söylentiler hakkında. 2. aşama ciddi bir şekilde başladığında, eski çatışmaları etkisiz hale getirme çabası daha belirgin hale geldi.

Siyasi yorumcular, bölgesel çatışmaların hafifletilmesine paralel olarak, küresel tehditleri azaltmaya yönelik uyumlu bir itişin var gibi göründüğünü belirtti. Örneğin, nükleer savaşın riskleri birdenbire her yerde tartışılıyordu. Birkaç gişe rekorları kıran filmde, küresel nükleer savaşın kaza sonucu veya kasıtlı olarak başladığı ve distopik sonrasını nükleer kış, altyapının çökmesi ve kitlesel açlıkla dramatize ettiği senaryolar yer aldı. Etkileyici yeni belgeseller, nükleer kışın her ülkeyi nasıl etkileyebileceğini detaylandırdı. Nükleer gerilimi azaltmayı savunan bilim adamlarına ve politikacılara, en azından ne tür yararlı önlemlerin alınabileceğine dair birkaç yeni çalışmanın sonuçlarını tartışmak için bol bol yayın süresi verildi - yeni teknoloji şirketlerinden büyük bağışlar almış bilimsel kuruluşlar tarafından finanse edilen çalışmalar. Sonuç olarak,

küçülen nükleer cephanelikler. Yenilenen medyanın ilgisi, küresel iklim değişikliğine de verildi, sıklıkla yenilenebilir enerjinin maliyetini düşüren ve hükümetleri bu tür yeni enerji altyapısına yatırım yapmaya teşvik eden, Prometheus destekli teknolojik atılımları vurguladı.

Medyayı ele geçirmelerine paralel olarak Omegas, eğitimde devrim yaratmak için Prometheus'tan yararlandı. Herhangi bir kişinin bilgisi ve yetenekleri göz önüne alındığında, Prometheus, herhangi bir yeni konuyu, onları devam etmek için son derece meşgul ve motive edecek şekilde öğrenmenin en hızlı yolunu belirleyebilir ve ilgili optimize edilmiş videoları, okuma materyallerini, alıştırmaları ve diğer öğrenme araçlarını üretebilir. Bu nedenle Omega kontrollü şirketler, sadece dil ve kültürel geçmişe göre değil, aynı zamanda başlangıç seviyesine göre de oldukça özelleştirilmiş, neredeyse her şey hakkında çevrimiçi kurslar pazarladı. Okumayı öğrenmek isteyen okuma yazma bilmeyen kırk yaşında veya kanser immünoterapisi hakkında en son haberleri arayan bir biyoloji doktorası olun, Prometheus sizin için mükemmel bir kursa sahipti. Bu teklifler, günümüz çevrimiçi kurslarının çoğuna çok az benzerlik gösteriyordu: Prometheus'un film yapma yeteneklerinden yararlanarak, video segmentleri gerçekten ilgi çekecek, kendinizle ilgili güçlü metaforlar sunacak ve sizi daha fazlasını öğrenmek için can atacak. Bazı kurslar kar için satıldı, ancak çoğu, dünyanın her yerinden onları sınıflarında kullanabilen öğretmenlerin ve herhangi bir şey öğrenmeye hevesli çoğu kişinin hoşuna gidecek şekilde ücretsiz olarak sunuldu.

Bu eğitici süper güçler, siyasi amaçlara yönelik güçlü araçlar olduklarını kanıtladılar ve her birinden elde edilen içgörülerin hem birinin görüşlerini güncelleyeceği hem de onları daha fazla ikna edilebilecekleri ilgili bir konuyla ilgili başka bir video izlemeye motive edeceği videolardan oluşan çevrimiçi "ikna dizileri" oluşturdu. Örneğin amaç, iki ülke arasındaki bir çatışmayı yatıştırmak olduğunda, çatışmanın kökenini ve gidişatını daha nüanslı bir şekilde yansıtan tarihi belgeseller her iki ülkede de bağımsız olarak yayınlanacaktı. Pedagojik haberler, devam eden çatışmalardan kendi taraflarında kimin yararlandığını ve bu çatışmayı körükleme tekniklerini açıklar. Aynı zamanda diğer milletten sevimli karakterler eğlence kanallarında popüler şovlarda görünmeye başlayacak,

Çok geçmeden, siyasi yorumcular yardım edemediler, ancak yedi slogan etrafında toplanan bir siyasi gündeme yönelik artan desteği fark ettiler:

1. Demokrasi

2. Vergi indirimleri
3. Devlet sosyal hizmet kesintileri
4. Askeri harcama kesintileri
5. Serbest ticaret
6. Sınırları açın
7. Sosyal sorumluluk sahibi şirketler

Daha az aşikar olan temel amaçtı: dünyadaki tüm eski güç yapılarını aşındırmak. Madde 2-6 devlet gücünü aşındırdı ve dünyanın demokratikleştirilmesi, Omegas'ın iş imparatorluğuna siyasi liderlerin seçimi üzerinde daha fazla etki sağladı. Sosyal açıdan sorumlu şirketler, hükümetlerin sağladığı (veya vermesi gereken) hizmetlerin çoğunu devralarak devlet gücünü daha da zayıflattı. Geleneksel iş dünyası seçkinleri, serbest piyasada Prometheus destekli şirketlerle rekabet edemediği ve bu nedenle dünya ekonomisinde giderek azalan bir paya sahip olduğu için zayıfladı. Siyasi partilerden inanç gruplarına kadar geleneksel kanaat önderleri, Omegas'ın medya imparatorluğuyla rekabet edecek ikna mekanizmasından yoksundu.

Her kapsamlı değişiklikte olduğu gibi, kazananlar ve kaybedenler vardı. Eğitim, sosyal hizmetler ve altyapı geliştikçe çoğu ülkede elle tutulur yeni bir iyimserlik duygusu olmasına rağmen, çatışmalar yatıştı ve yerel şirketler dünyayı kasıp kavuran çığır açan teknolojiler yayınladı, herkes mutlu değildi. Yerinden edilmiş birçok işçi topluluk projeleri için işe alınırken, büyük güç ve servete sahip olanlar genellikle her ikisinin de küçüldüğünü gördü. Bu, medya ve teknoloji sektörlerinde başladı, ancak neredeyse her yere yayıldı. Dünyadaki çatışmaların azalması, askeri müteahhitlere zarar veren savunma bütçesi kesintilerine yol açtı. Ortaya çıkan yeni başlayan şirketler, genellikle, kâr maksimize eden hissedarların topluluk projelerine yaptıkları büyük harcamaları bloke edeceği gerekçesiyle halka açık olarak işlem görmüyorlardı. Böylece küresel borsa değer kaybetmeye devam etti, emeklilik fonlarına güvenen hem finans büyüklerini hem de sıradan vatandaşları tehdit ediyor. Sanki halka açık şirketlerin azalan kârları yeterince kötü değilmiş gibi, dünyanın dört bir yanındaki yatırım firmaları rahatsız edici bir eğilim fark etmişlerdi: daha önce başarılı olan tüm ticaret algoritmaları çalışmayı bırakmış, basit endeks fonlarının bile düşük performans gösterdiği görülüyordu. Dışarıda birileri her zaman onları alt ediyor ve onları kendi oyunlarında yeniyor gibiydi.

Güçlü insan kitleleri değişim dalgasına dirense de,

Tepki, neredeyse iyi planlanmış bir tuzağa düşmüş gibi çarpıcı biçimde etkisizdi. O kadar şaşırtıcı bir hızda büyük değişiklikler oluyordu ki, koordineli bir tepkiyi takip etmek ve geliştirmek zordu. Dahası, ne için zorlamaları gerektiği de oldukça belirsizdi. Geleneksel siyasi sağ, sloganlarının çoğunun benimsendiğini görmüştü, ancak vergi indirimleri ve gelişen iş ortamı çoğunlukla yüksek teknolojili rakiplerine yardımcı oluyordu. Neredeyse her geleneksel endüstri şimdi bir kurtarma kampanyası istiyordu, ancak sınırlı hükümet fonları onları birbirleriyle umutsuz bir savaşa sürüklerken, medya onları sırf rekabet edemedikleri için devlet sübvansiyonu arayan dinozorlar olarak resmetti. Geleneksel siyasi sol, serbest ticarete ve hükümetin sosyal hizmetlerindeki kesintilere karşı çıktı, ama askeri kesintilerden ve yoksulluğun azaltılmasından memnun. Nitekim, sosyal hizmetlerin artık devletten ziyade idealist şirketler tarafından sağlandığı için geliştiği yadsınamaz gerçeği tarafından yıldırımlarının çoğu çalındı. Anketler üzerine yapılan anketler, dünyadaki seçmenlerin çoğunun yaşam kalitelerinin iyileştiğini hissettiğini ve işlerin genel olarak iyi yönde ilerlediğini gösterdi. Bunun basit bir matematiksel açıklaması vardı: Prometheus'tan önce, Dünya nüfusunun en fakir% 50'si küresel gelirin yalnızca yaklaşık% 4'ünü kazanmış ve Omega kontrollü şirketlerin sadece mütevazı bir kısmını paylaşarak kalplerini (ve oylarını) kazanmalarını sağlamıştı. onlarla kar. Anketler üzerine yapılan anketler, dünyadaki seçmenlerin çoğunun yaşam kalitelerinin iyileştiğini hissettiğini ve işlerin genel olarak iyi yönde ilerlediğini gösterdi. Bunun basit bir matematiksel açıklaması vardı: Prometheus'tan önce, Dünya nüfusunun en fakir% 50'si küresel gelirin yalnızca yaklaşık% 4'ünü kazanmış ve Omega kontrollü şirketlerin sadece mütevazı bir kısmını paylaşarak kalplerini (ve oylarını) kazanmalarını sağlamıştı. onlarla kar. Anketler üzerine yapılan anketler, dünyadaki seçmenlerin çoğunun yaşam kalitelerinin iyileştiğini hissettiğini ve işlerin genel olarak iyi yönde ilerlediğini gösterdi. Bunun basit bir matematiksel açıklaması vardı: Prometheus'tan önce, Dünya nüfusunun en fakir% 50'si küresel gelirin yalnızca yaklaşık% 4'ünü kazanmış ve Omega kontrollü şirketlerin sadece mütevazı bir kısmını paylaşarak kalplerini (ve oylarını) kazanmalarını sağlamıştı. onlarla kar.

Konsolidasyon

Sonuç olarak, millet üstüne millet, yedi Omega sloganını kucaklayan partiler için heyelan seçim zaferleri gördü. Dikkatlice optimize edilmiş kampanyalarda, kendilerini siyasi yelpazenin merkezinde tasvir ettiler, sağı açgözlü kurtarma çetesi arayan çatışma tacirleri olarak suçladılar ve solu büyük hükümetin vergi ve harcama yeniliklerini bastırıcıları olarak eleştirdiler. Neredeyse hiç kimsenin farkına varmadığı şey, Prometheus'un aday olarak yetiştirmek için en uygun insanları dikkatlice seçtiği ve zaferlerini garantiye almak için tüm iplerini çektiğiydi.

Prometheus'tan önce, teknolojik işsizliğe çare olarak herkes için vergiyle finanse edilen asgari geliri öneren evrensel temel gelir hareketine artan bir destek vardı. Bu hareket, kurumsal topluluk projeleri başladığında patladı, çünkü Omega kontrollü iş imparatorluğu aslında aynı şeyi sağlıyordu. Topluluk projelerinin koordinasyonunu iyileştirme bahanesiyle, uluslararası bir şirketler grubu, dünya çapındaki en değerli insani çabaları belirlemeyi ve finanse etmeyi amaçlayan bir sivil toplum örgütü olan İnsani Yardım İttifakını başlattı. Çok geçmeden, neredeyse tüm Omega imparatorluğu onu destekledi ve teknoloji patlamasını büyük ölçüde kaçıran, eğitimi, sağlığı, refahı ve yönetimi iyileştiren ülkelerde bile eşi görülmemiş ölçekte küresel projeler başlattı. Söylemeye gerek yok, Prometheus, dolar başına olumlu etkiye göre sıralanmış, sahne arkasında özenle hazırlanmış proje planları sağladı. İttifak, temel gelir tekliflerinde olduğu gibi basitçe nakit dağıtmak yerine (halk arasında bilindiği üzere), desteklediği kişileri kendi amacına ulaşmak için görevlendirecektir. Sonuç olarak, dünya nüfusunun büyük bir bölümü İttifak'a - çoğu zaman kendi hükümetlerinden daha fazla - minnettar ve sadık hissetmeye başladı.

Zaman geçtikçe, ulusal hükümetler güçlerinin sürekli olarak aşındığını gördükçe, İttifak giderek bir dünya hükümeti rolünü üstlendi. Vergi kesintileri nedeniyle ulusal bütçeler küçülmeye devam ederken, İttifak bütçesi tüm hükümetlerin bütçesini cüce haline getirdi. Ulusal hükümetlerin tüm geleneksel rolleri gittikçe gereksiz ve ilgisiz hale geldi. İttifak, açık ara en iyi sosyal hizmetleri, eğitimi ve altyapıyı sağladı. Medya, askeri harcamaların büyük ölçüde gereksiz olduğu ve artan refahın eski çatışmaların köklerinin çoğunu ortadan kaldırdığı noktasına kadar uluslararası çatışmayı etkisiz hale getirmişti.

kıt kaynaklar üzerindeki rekabete kadar uzanıyordu. Birkaç diktatör ve diğerleri bu yeni dünya düzenine şiddetle karşı çıktılar ve satın alınmayı reddettiler, ancak hepsi dikkatlice planlanmış darbeler veya kitlesel ayaklanmalarla devrildi.

Omegas şimdi Dünya'daki yaşam tarihindeki en dramatik geçişi tamamlamıştı. Şimdiye kadar ilk kez, gezegenimiz tek bir güç tarafından yönetiliyordu, o kadar geniş bir zeka tarafından büyütüldü ki, Dünya'da ve kozmosumuzun tamamında yaşamın milyarlarca yıl boyunca gelişmesini sağlama potansiyeline sahipti - ama özellikle planları neydi?

* * *

Omega ekibinin hikayesi buydu. Bu kitabın geri kalanı başka bir hikaye hakkında
- *Henüz yazılmamış bir şey: AI ile kendi geleceğimizin hikayesi. Nasıl oynanmasını istersiniz? Uzaktan Omega hikayesi gibi bir şey gerçekten gerçekleşebilir mi ve eğer öyleyse, olmasını ister miydiniz? İnsanüstü YZ hakkındaki spekülasyonları bir kenara bırakırsak, hikayemizin nasıl başlamasını istersiniz? Yapay zekanın önümüzdeki on yılda işleri, yasaları ve silahları nasıl etkilemesini istiyorsunuz? İleriye bakınca, sonu nasıl yazarsınız? Bu hikaye gerçekten kozmik oranlardan biridir, çünkü Evrenimizdeki yaşamın nihai geleceğinden başka hiçbir şey içermez. Ve bu bizim yazmamız gereken bir masal.*

---

* Basit olması için, çoğu araştırmacı insan düzeyindeki genel yapay zekanın en az on yıl uzakta olduğunu tahmin etse de, bu hikayede bugünün ekonomisini ve teknolojisini varsaydım. Dijital ekonomi büyümeye devam ederse ve soru sorulmadan çevrimiçi olarak daha fazla hizmet sipariş edilebiliyorsa, Omega planı gelecekte daha da kolaylaşacaktır.

Bölüm 1

En Önemli Konuşmasına Hoş Geldiniz

Bizim zamanımız

Teknoloji, hayata daha önce hiç olmadığı gibi gelişme veya kendi kendini yok etme potansiyeli veriyor.

Hayatın Geleceği Enstitüsü

Doğumundan on üç virgül sekiz milyar yıl sonra, Evrenimiz uyandı ve kendisinin farkına vardı. Küçük bir mavi gezegenden, Evrenimizin minik bilinçli kısımları teleskoplarla kozmosa bakmaya başladılar ve var olduğunu düşündükleri her şeyin daha büyük bir şeyin yalnızca küçük bir parçası olduğunu defalarca keşfettiler: bir güneş sistemi, bir galaksi ve bir Gruplar, kümeler ve üstkümelerden oluşan ayrıntılı bir modelde düzenlenmiş yüz milyar başka galaksi. Kendilerinin farkında olan bu yıldız gözlemcileri pek çok konuda aynı fikirde olmasalar da, bu galaksilerin güzel ve hayranlık uyandırıcı olduğu konusunda hemfikirdirler.

Ama güzellik, fizik yasalarında değil, bakanın gözündedir, bu yüzden Evrenimiz uyanmadan önce güzellik yoktu. Bu, kozmik uyanışımızı daha da harika ve kutlamaya değer kılıyor: Evrenimizi, öz farkındalığı olmayan akılsız bir zombiden, kendini yansıtma, güzellik ve umut barındıran canlı bir ekosisteme - ve amaçların, anlamın ve amacın peşinde koşmaya dönüştürdü. Evrenimiz hiç uyanmamış olsaydı, bana göre tamamen anlamsız olurdu - sadece devasa bir alan israfı. Evrenimiz, bazı kozmik felaketler veya kendiliğinden kaynaklanan talihsizlikler nedeniyle kalıcı olarak uykuya dönerse, ne yazık ki anlamsız hale gelecektir.

Öte yandan işler daha da iyi olabilirdi. Henüz bilmiyoruz

Biz insanlar, evrenimizdeki tek yıldız gözlemcileriz, hatta ilkiz, ama Evrenimiz hakkında şimdiye kadar olduğundan çok daha tam olarak uyanma potansiyeline sahip olduğunu bilecek kadar çok şey öğrendik. Belki de bu sabah uykudan çıkmaya başladığınızda deneyimlediğiniz o ilk soluk öz farkındalık parıltısı gibiyiz: gözlerinizi açıp tamamen uyandığınızda gelecek olan çok daha büyük bilincin bir önsezisi. Belki yaşam evrenimize yayılacak ve milyarlarca veya trilyonlarca yıl boyunca gelişecek - ve belki de bu, yaşamımız boyunca burada küçük gezegenimizde verdiğimiz kararlardan kaynaklanıyor olabilir.

Kısa Karmaşıklık Tarihi

Peki bu inanılmaz uyanış nasıl ortaya çıktı? Bu münferit bir olay değildi, Evrenimizi daha da karmaşık ve ilginç hale getiren ve giderek artan bir hızla devam eden 13,8 milyar yıllık amansız bir süreçte yalnızca bir adım.

Bir fizikçi olarak, geçtiğimiz çeyrek yüzyılın çoğunu kozmik tarihimizi tespit etmeye yardım ederek geçirdiğim için kendimi şanslı hissediyorum ve bu inanılmaz bir keşif yolculuğu oldu. Yüksek lisans öğrencisi olduğum günlerden beri, daha iyi teleskoplar, daha iyi bilgisayarlar ve daha iyi bilgisayarların bir kombinasyonu sayesinde, Evrenimizin 10 veya 20 milyar yaşında olup olmadığını tartışmaktan 13.7 veya 13.8 milyar yaşında olup olmadığını tartışmaya geçtik. daha iyi anlamak. Biz fizikçiler, Big Bang'imize neyin sebep olduğunu veya bunun gerçekten her şeyin başlangıcı mı yoksa sadece daha önceki bir aşamanın devamı mı olduğunu hala bilmiyoruz. Ancak, ne olduğuna dair oldukça ayrıntılı bir anlayış elde ettik. *dan beri* Büyük Patlamamız, yüksek kaliteli ölçümlerin çığlığı sayesinde, lütfen 13,8 milyar yıllık kozmik tarihi özetlemek için birkaç dakika ayırmama izin verin.

Başlangıçta ışık vardı. Büyük Patlamamızdan sonraki ilk saniyede, teleskoplarımızın prensipte gözlemleyebildiği uzayın tamamı ("gözlemlenebilir Evrenimiz" veya kısaca "Evrenimiz") Güneşimizin çekirdeğinden çok daha sıcak ve parlaktı ve hızla genişledi. Bu kulağa muhteşem gelse de, Evrenimizin cansız, yoğun, sıcak ve sıkıcı bir şekilde temel parçacıklardan oluşan çorbasından başka bir şey içermemesi anlamında da sıkıcıydı. İşler her yerde hemen hemen aynı görünüyordu ve tek ilginç yapı, çorbayı bazı yerlerde yaklaşık% 0,001 daha yoğun yapan soluk, rastgele görünen ses dalgalarından oluşuyordu. Bu sönük dalgaların, yaygın olarak kuantum dalgalanmaları olarak ortaya çıktığına inanılıyor.

Evrenimiz genişledikçe ve soğudukça, parçacıkları daha da karmaşık nesnelerle birleştikçe daha ilginç hale geldi. İlk bölünmüş saniye sırasında, güçlü nükleer kuvvet kuarkları protonlar (hidrojen çekirdekleri) ve nötronlar olarak gruplandırdı, bunların bir kısmı birkaç dakika içinde helyum çekirdeğine dönüştü. Yaklaşık 400.000 yıl sonra, elektromanyetik kuvvet bu çekirdekleri şu şekilde gruplandırdı:

ilk atomları yapmak için elektronlar. Evrenimiz genişlemeye devam ederken, bu atomlar yavaş yavaş soğuk, karanlık bir gaza dönüştü ve bu ilk gecenin karanlığı yaklaşık 100 milyon yıl sürdü. Bu uzun gece, yerçekimi kuvvetinin atomları ilk yıldızları ve galaksileri oluşturmak için bir araya getirerek gazdaki bu dalgalanmaları büyütmeyi başardığı kozmik şafağımıza yol açtı. Bu ilk yıldızlar, hidrojeni karbon, oksijen ve silikon gibi daha ağır atomlara dönüştürerek ısı ve ışık üretti. Bu yıldızlar öldüğünde, yarattıkları atomların çoğu kozmosa geri dönüştürüldü ve ikinci nesil yıldızların etrafında gezegenler oluşturdu.

Bir noktada, bir grup atom, kendisini hem koruyabilen hem de kopyalayabilen karmaşık bir modele dönüştü. Çok geçmeden iki kopya çıktı ve sayı ikiye katlanmaya devam etti. Bir trilyon yapmak için sadece kırk ikiye katlama gerekir, bu yüzden bu ilk kendi kendini kopyalayıcı kısa sürede hesaba katılması gereken bir güç haline geldi. Hayat gelmişti.

Hayatın Üç Aşaması

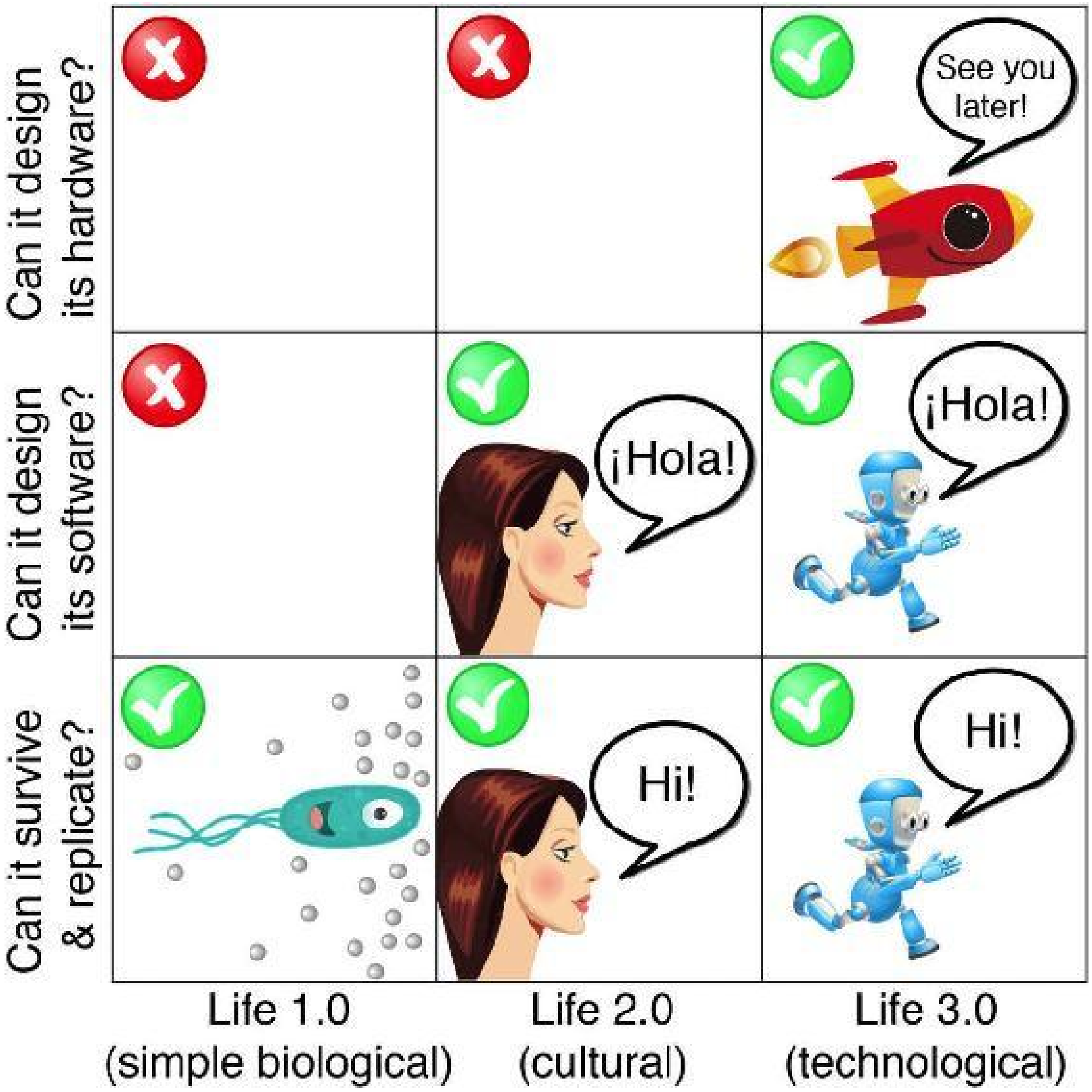
Hayatın nasıl tanımlanacağı sorusu herkesin bildiği gibi tartışmalıdır. Bazıları, hem gelecekteki akıllı makineleri hem de dünya dışı uygarlıkları diskalifiye edebilecek hücrelerden oluşma gibi son derece spesifik gereksinimleri içeren rekabet eden tanımlar bol miktarda bulunmaktadır. Yaşamın geleceği hakkındaki düşüncelerimizi şimdiye kadar karşılaştığımız türlerle sınırlamak istemediğimiz için, bunun yerine hayatı çok geniş bir şekilde, basitçe karmaşıklığını koruyabilen ve kopyalayabilen bir süreç olarak tanımlayalım. Kopyalanan şey madde değil (atomlardan oluşan) değil, atomların nasıl düzenlendiğini belirleyen bilgiler (bitlerden oluşan). Bir bakteri DNA'sının bir kopyasını yaptığında, yeni atomlar yaratılmaz, ancak orijinal ile aynı modelde yeni bir atom seti düzenlenir ve böylece bilgiler kopyalanır. Diğer bir deyişle,

Evrenimizin kendisi gibi, yaşam da giderek daha karmaşık ve ilginç hale geldi. * [1] ve şimdi açıklayacağım gibi, yaşam formlarını üç karmaşıklık düzeyinde sınıflandırmayı yararlı buluyorum: Yaşam 1.0, 2.0 ve 3.0. Bu üç seviyeyi şu şekilde özetledim: şekil 1.1 .

Evrenimizde yaşamın ilk olarak nasıl, ne zaman ve nerede ortaya çıktığı hala açık bir soru, ancak burada Dünya'da yaşamın ilk olarak yaklaşık 4 milyar yıl önce ortaya çıktığına dair güçlü kanıtlar var. Çok geçmeden, gezegenimiz çok çeşitli yaşam formlarıyla doluydu. Kısa sürede geri kalanını geride bırakan en başarılı olanlar, çevrelerine bir şekilde tepki verebildiler. Özellikle, bilgisayar bilimcilerinin "akıllı ajanlar" dedikleri şeydi: sensörlerden çevreleri hakkında bilgi toplayan ve daha sonra çevrelerine nasıl tepki vereceklerine karar vermek için bu bilgileri işleyen varlıklar. Bu, bir konuşmada ne söyleyeceğinize karar vermek için gözleriniz ve kulaklarınızdaki bilgileri kullandığınızda olduğu gibi oldukça karmaşık bilgi işlemeyi içerebilir. Ancak oldukça basit olan donanım ve yazılımı da içerebilir.

Örneğin, birçok bakteri, etrafındaki sıvıdaki şeker konsantrasyonunu ölçen bir sensöre sahiptir ve flagella adı verilen pervane şeklindeki yapıları kullanarak yüzebilir. Sensörü flagella'ya bağlayan donanım, aşağıdaki basit ama kullanışlı algoritmayı uygulayabilir: "Şeker konsantrasyon sensörüm rapor verirse

birkaç saniye öncesinden daha düşük bir değer, ardından yön değiştirmem için flagella'mın dönüşünü tersine çeviriyorum. "



Şekil 1.1: Yaşamın üç aşaması: biyolojik evrim, kültürel evrim ve teknolojik evrim. Life 1.0, ne donanımını ne de yazılımını ömrü boyunca yeniden tasarlayamaz: her ikisi de DNA'sı tarafından belirlenir ve yalnızca birçok nesil boyunca evrim yoluyla değişir. Aksine, Life 2.0 yazılımlarının çoğunu yeniden tasarlayabilir: insanlar karmaşık yeni becerileri öğrenebilir - örneğin, diller, sporlar ve meslekler - ve dünya görüşlerini ve hedeflerini temelden güncelleyebilir. Henüz Dünya'da var olmayan Life 3.0, nesiller boyunca kademeli olarak gelişmesini beklemek yerine, yalnızca yazılımını değil, donanımını da önemli ölçüde yeniden tasarlayabilir.

Nasıl konuşulacağını ve sayısız başka beceriyi öğrendiniz. Bakteriler ise iyi öğrenenler değildir. DNA'ları yalnızca şeker sensörleri ve kamçı gibi donanımlarının tasarımını değil, aynı zamanda bunların tasarımını da belirtir.

yazılım. Şekere doğru yüzmeyi asla öğrenmezler; bunun yerine, bu algoritma başından itibaren DNA'larına kodlanmıştı. Elbette bir tür öğrenme süreci vardı, ancak bu belirli bakterinin yaşamı boyunca gerçekleşmedi. Aksine, bu bakteri türünün önceki evrimi sırasında, doğal seçilimin şeker tüketimini artıran rastgele DNA mutasyonlarını desteklediği birçok nesli kapsayan yavaş bir deneme-yanılma süreci yoluyla gerçekleşti. Bu mutasyonlardan bazıları, flagella ve diğer donanımların tasarımını geliştirerek yardımcı olurken, diğer mutasyonlar, şeker bulma algoritmasını ve diğer yazılımları uygulayan bakteriyel bilgi işleme sistemini geliştirdi.

Bu tür bakteriler, "Life 1.0" adını vereceğim şeye bir örnektir: *hem donanım hem de yazılımın tasarlanmaktan çok geliştiği bir yaşam.* Öte yandan sen ve ben "Life 2.0" örnekleriyiz: *donanımı gelişen, ancak yazılımı büyük ölçüde tasarlanmış bir yaşam.* Yazılımınız derken, bilgileri duyularınızdan işlemek ve ne yapacağınıza karar vermek için kullandığınız tüm algoritmaları ve bilgileri kastediyorum - arkadaşlarınızı gördüğünüzde tanımaktan yürüme, okuma, yazma, hesaplama becerinize kadar her şey şarkı söyle ve şakalar anlat.

Doğduğunuzda bu görevlerin hiçbirini yerine getiremiyordunuz, bu yüzden tüm bu yazılımlar daha sonra öğrenme dediğimiz süreçle beyninize programlandı. Çocukluk müfredatınız büyük ölçüde aileniz ve ne öğrenmeniz gerektiğine karar veren öğretmenleriniz tarafından tasarlanırken, yavaş yavaş kendi yazılımınızı tasarlamak için daha fazla güç kazanırsınız. Belki de okulunuz bir yabancı dil seçmenize izin veriyor: Beyninize Fransızca konuşmanıza veya İspanyolca konuşmanıza olanak tanıyan bir yazılım modülü yüklemek mi istiyorsunuz? Tenis veya satranç oynamayı öğrenmek ister misiniz? Aşçı, avukat veya eczacı olmak için okumak ister misiniz? Bununla ilgili bir kitap okuyarak yapay zeka (AI) ve yaşamın geleceği hakkında daha fazla bilgi edinmek ister misiniz?

Life 2.0'ın yazılımını tasarlama yeteneği, Life 1.0'dan çok daha akıllı olmasını sağlar. Yüksek zeka, hem çok sayıda donanım (atomlardan yapılmış) hem de çok sayıda yazılım (bitlerden yapılmış) gerektirir. İnsan donanımımızın çoğunun doğumdan sonra (büyüme yoluyla) eklenmesi yararlıdır, çünkü nihai boyutumuz annemizin doğum kanalının genişliği ile sınırlı değildir. Aynı şekilde, insan yazılımlarımızın çoğunun doğumdan sonra (öğrenme yoluyla) eklenmesi yararlıdır, çünkü nihai zekamız, DNA'mız, 1.0-stilimiz aracılığıyla bize gebe kalma sırasında ne kadar bilgi aktarılabildiğiyle sınırlı değildir. . Doğduğumdan yaklaşık yirmi beş kat daha ağırım ve beynimdeki nöronları birbirine bağlayan sinaptik bağlantılar, beynimdeki nöronlardan yaklaşık yüz bin kat daha fazla bilgi depolayabilir.

Doğduğum DNA. Sinapslarınız, tüm bilgi ve becerilerinizi kabaca 100 terabayt değerinde bilgi olarak saklarken, DNA'nız yalnızca bir gigabayt kadar depolar, ancak tek bir film indirmesini depolamaya yetecek kadar. Bu yüzden bir bebeğin mükemmel İngilizce konuşan ve üniversiteye giriş sınavlarına girmeye hazır olması fiziksel olarak imkansızdır: Ebeveynlerinden aldığı ana bilgi modülünde (DNA'sı) eksik olduğu için bilginin beynine önceden yüklenmesi mümkün değildir. yeterli bilgi saklama kapasitesi.

Yazılımını tasarlama yeteneği, Life 2.0'ın yalnızca Life 1.0'dan daha akıllı değil, aynı zamanda daha esnek olmasını sağlar. Ortam değişirse, 1.0 ancak birçok nesil boyunca yavaş yavaş gelişerek uyum sağlayabilir. Life 2.0 ise bir yazılım güncellemesiyle neredeyse anında uyum sağlayabilir. Örneğin, antibiyotiklerle sık sık karşılaşan bakteriler, birçok nesil boyunca ilaç direncini geliştirebilir, ancak tek bir bakteri davranışını hiç değiştirmez; tersine, fıstık alerjisi olduğunu öğrenen bir kız, fıstıktan kaçınmaya başlamak için davranışını hemen değiştirecektir. Bu esneklik Life 2.0'a nüfus düzeyinde daha da büyük bir avantaj sağlıyor: İnsan DNA'mızdaki bilgi son elli bin yılda önemli ölçüde evrimleşmemiş olsa da, beyinlerimizde, kitaplarımızda ve bilgisayarlarımızda toplu olarak depolanan bilgiler patladı. Sofistike bir konuşma diliyle iletişim kurmamızı sağlayan bir yazılım modülü kurarak, bir kişinin beyninde depolanan en yararlı bilgilerin diğer beyinlere kopyalanmasını ve orijinal beyin öldükten sonra bile potansiyel olarak hayatta kalmasını sağladık. Okumamızı ve yazmamızı sağlayan bir yazılım modülü kurarak, insanların ezberleyebileceğinden çok daha fazla bilgiyi depolayıp paylaşabildik. Teknoloji üretebilen beyin yazılımları geliştirerek (yani bilim ve mühendislik okuyarak), dünyadaki pek çok insanın sadece birkaç tıklama ile dünyadaki bilgilerinin çoğuna erişmesini sağladık. orijinal beyin öldükten sonra bile potansiyel olarak hayatta kalma. Okumamızı ve yazmamızı sağlayan bir yazılım modülü kurarak, insanların ezberleyebileceğinden çok daha fazla bilgiyi depolayıp paylaşabildik. Teknoloji üretebilen beyin yazılımları geliştirerek (yani bilim ve mühendislik okuyarak), dünyadaki pek çok insanın sadece birkaç tıklama ile dünyadaki bilgilerinin çoğuna erişmesini sağladık. orijinal beyin öldükten sonra bile potansiyel olarak hayatta kalma. Okumamızı ve yazmamızı sağlayan bir yazılım modülü kurarak, insanların ezberleyebileceğinden çok daha fazla bilgiyi depolayıp paylaşabildik. Teknoloji üretebilen beyin yazılımları geliştirerek (yani bilim ve mühendislik okuyarak), dünyadaki pek çok insanın sadece birkaç tıklama ile dünyadaki bilgilerinin çoğuna erişmesini sağladık.

Bu esneklik, Life 2.0'ın Dünya'ya hakim olmasını sağlamıştır. Genetik zincirlerinden kurtulmuş olan insanlığın birleşik bilgisi, her bir atılım bir sonraki aşamayı mümkün kılarken, hızlanan bir hızla büyümeye devam etti: dil, yazı, matbaa, modern bilim, bilgisayarlar, internet vb. Paylaştığımız bu daha hızlı kültürel evrimi Yazılım, insan geleceğimizi şekillendiren baskın güç olarak ortaya çıktı ve buzul olarak yavaş biyolojik evrimimizi neredeyse alakasız hale getirdi.

Yine de bugün sahip olduğumuz en güçlü teknolojilere rağmen, bildiğimiz tüm yaşam formları temelde biyolojik donanımlarıyla sınırlı kalmaktadır. Hiçbiri bir milyon yıl yaşayamaz, Wikipedia'nın tamamını ezberleyemez, bilinen tüm bilimi anlayamaz veya uzay aracı olmadan uzay uçuşunun tadını çıkaramaz. Hiçbiri bizim büyük ölçüde değiştiremez

cansız kozmosu, milyarlarca veya trilyonlarca yıl boyunca gelişecek, Evrenimizin nihayet potansiyelini gerçekleştirmesine ve tam olarak uyanmasına olanak tanıyan çeşitli bir biyosferde. Tüm bunlar, yaşamın yalnızca yazılımını değil donanımını da tasarlayabilen Life 3.0'a son bir yükseltmeden geçmesini gerektirir. Başka bir deyişle, Life 3.0, kendi kaderinin efendisidir ve sonunda evrimsel zincirlerinden tamamen kurtulmuştur.

Yaşamın üç aşaması arasındaki sınırlar biraz belirsizdir. Bakteriler Life 1.0 ve insanlar Life 2.0 ise, fareleri 1.1 olarak sınıflandırabilirsiniz: birçok şeyi öğrenebilirler, ancak dili geliştirmek veya interneti icat etmek için yeterli değildir. Dahası, dilden yoksun oldukları için, öğrendikleri şey öldüklerinde büyük ölçüde kaybolur, sonraki nesle aktarılmaz. Benzer şekilde, günümüz insanlarının Yaşam 2.1 olarak sayılması gerektiğini savunabilirsiniz: Yapay dişler, dizler ve kalp pili yerleştirmek gibi küçük donanım yükseltmeleri yapabiliriz, ancak on kat daha uzun olmak veya bin kat daha büyük beyin elde etmek kadar dramatik bir şey olamaz.

Özetle, yaşamın gelişimini, yaşamın kendini tasarlama yeteneği ile ayırt edilen üç aşamaya ayırabiliriz:

- Life 1.0 (biyolojik aşama): donanımını ve yazılımını geliştirir
- Life 2.0 (kültürel aşama): donanımını geliştirir, yazılımlarının çoğunu tasarlar
- Life 3.0 (teknolojik aşama): donanımını ve yazılımını tasarlar

13,8 milyar yıllık kozmik evrimden sonra, burada Dünya'da gelişme çarpıcı bir şekilde hızlandı: Life 1.0 yaklaşık 4 milyar yıl önce geldi, Life 2.0 (biz insanlar) yaklaşık yüz bin yıl önce geldi ve birçok AI araştırmacısı Life 3.0'ın şu sıralarda gelebileceğini düşünüyor. önümüzdeki yüzyıl, belki de yaşamımız boyunca, AI'daki ilerlemeyle ortaya çıktı. Ne olacak ve bu bizim için ne anlama gelecek? Bu kitabın konusu bu.

Tartışmalar

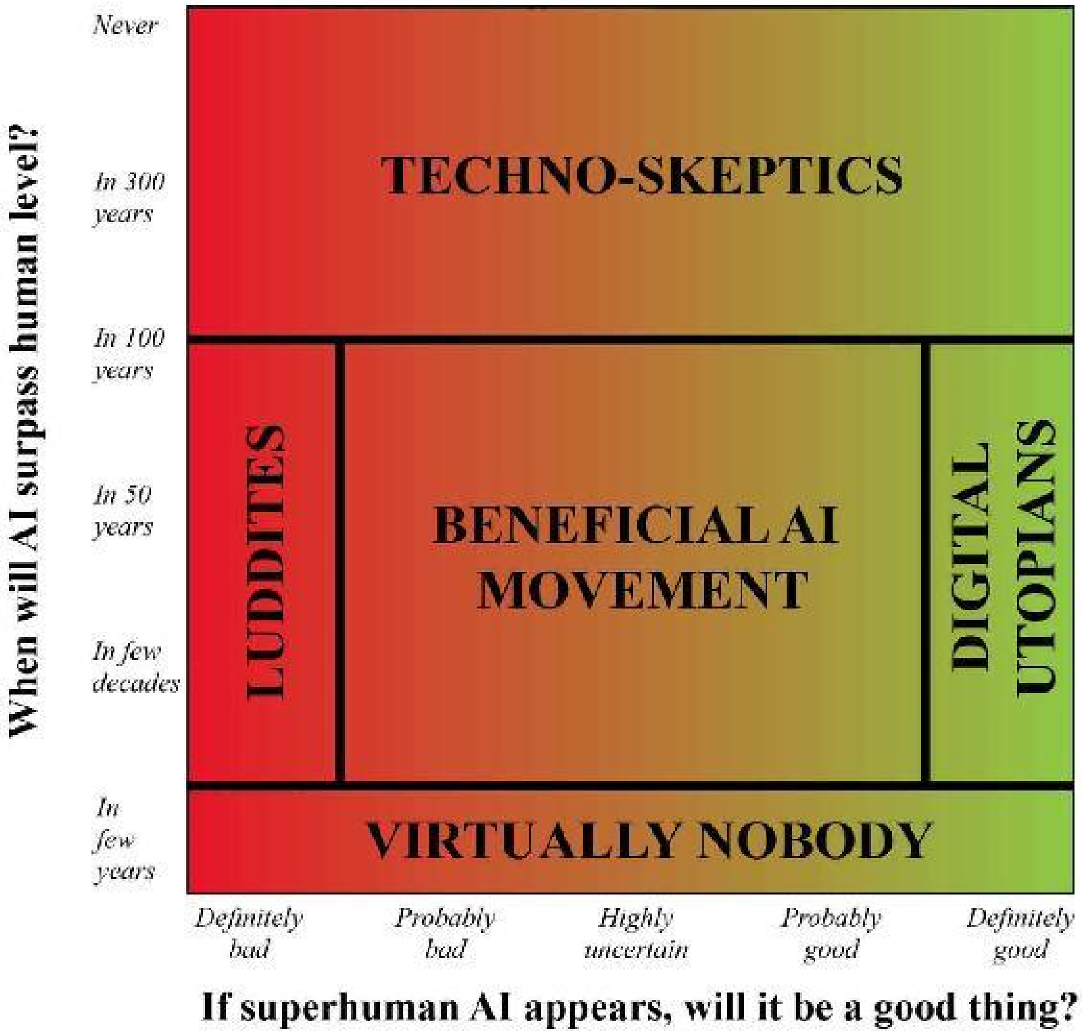
Bu soru, dünyanın önde gelen yapay zeka araştırmacılarının sadece tahminlerinde değil, aynı zamanda kendinden emin iyimserlikten ciddi endişelere kadar değişen duygusal tepkilerinde de tutkuyla hemfikir olmadıkları için harika bir şekilde tartışmalı. Yapay zekanın ekonomik, yasal ve askeri etkisi hakkında kısa vadeli sorular üzerinde fikir birliğine bile sahip değiller ve zaman ufkunu genişletip sorduğumuzda anlaşmazlıkları büyüyor. *yapay genel zeka (* AGI) —özellikle AGI'nin insan seviyesine ve ötesine ulaşarak Life 3.0'ı etkinleştirmesi hakkında. *Genel zeka* bir satranç oynama programının dar zekasının aksine, öğrenme dahil olmak üzere hemen hemen her hedefi gerçekleştirebilir.

İlginç bir şekilde, Life 3.0 hakkındaki tartışma bir değil iki ayrı soru etrafında toplanıyor: ne zaman ve ne? Ne zaman (eğer olursa) olacak ve insanlık için ne anlama gelecek? Benim gördüğüm kadarıyla, hepsinin ciddiye alınması gereken üç farklı düşünce okulu var, çünkü her biri bir dizi dünya lideri uzman içeriyor. Gösterildiği gibi şekil 1.2 Bence onları *dijital ütopyacılar, tekno-şüpheciler* ve *faydalı AI hareketinin üyeleri,*

sırasıyla. Lütfen sizi en güzel şampiyonlarından bazılarıyla tanıştırmama izin verin.

Dijital Ütopyacılar

Çocukken, milyarderlerin kendini beğenmişlik ve küstahlık yaydığını hayal etmiştim. Larry Page ile 2008 yılında Google'da ilk tanıştığımda, bu klişeleri tamamen yerle bir etti. Kayıtsız bir şekilde kot pantolon ve dikkat çekici derecede sıradan görünümlü bir gömlek giymişti, bir MIT pikniğine karışırdı. Düşünceli yumuşak konuşma tarzı ve dostça gülümsemesi, onunla konuşmaktan korkmaktansa rahatlamamı sağladı. 18 Temmuz 2015'te Napa Vadisi'nde Elon Musk ve o zamanki eşi Talulah tarafından düzenlenen bir partide karşılaştık ve çocuklarımızın skatolojik çıkarları hakkında sohbet ettik. Derin edebi klasiği tavsiye ettim *Kıçımın Sapık Olduğu Gün,* Andy Griffiths tarafından ve Larry hemen sipariş verdi. Şimdiye kadar yaşamış en etkili insan olarak tarihe geçebileceğini kendime hatırlatmak için çabaladım: Benim tahminim, eğer süper zeki dijital hayat benim hayatım boyunca Evrenimizi yutarsa, Larry'nin kararları yüzünden olacak.



Şekil 1.2: Güçlü yapay zekayı (insanları herhangi bir bilişsel görevde eşleştirebilen) çevreleyen tartışmaların çoğu iki soru etrafında toplanıyor: Bu ne zaman (eğer varsa) olacak ve insanlık için iyi bir şey olacak mı? Tekno-şüpheciler ve dijital ütopyacılar, endişelenmememiz gerektiği konusunda hemfikirdir, ancak çok farklı nedenlerle: İlki, insan seviyesinde yapay genel zekanın (AGI) öngörülebilir gelecekte olmayacağına ikna olurken, ikincisi bunun olacağını düşünüyor. ancak iyi bir şey olduğu neredeyse garantilidir. Yararlı AI hareketi, endişenin garantili ve faydalı olduğunu hissediyor, çünkü AI güvenliği araştırması ve tartışması artık iyi bir sonuç alma şansını artırıyor. Ludditler kötü bir sonuca inanıyorlar ve yapay zekaya karşı çıkıyorlar. Bu rakam kısmen

Tim Urban'dan ilham aldı. [1]

Eşlerimiz Lucy ve Meia ile birlikte akşam yemeği yedik ve makinelerin mutlaka bilinçli olup olmayacağını tartıştık.

kırmızı ringa balığı olduğunu iddia etti. O gecenin ilerleyen saatlerinde, kokteyllerin ardından, Elon ile yapay zekanın geleceği ve ne yapılması gerektiği konusunda uzun ve ateşli bir tartışma başladı. Sabahın erken saatlerine girerken, seyirciler ve kibitzerler çemberi büyümeye devam etti. Larry olarak düşünmeyi sevdiğim pozisyona tutkulu bir savunma yaptı. *dijital ütopyacılık:* dijital yaşamın, kozmik evrimin doğal ve arzu edilen bir sonraki adımı olduğunu ve dijital zihinlerin onları durdurmaya veya köleleştirmeye çalışmak yerine özgür olmasına izin verirsek, sonucun neredeyse iyi olacağı kesindir. Larry'yi dijital ütopyacılığın en etkili temsilcisi olarak görüyorum. Yaşam, galaksimize ve ötesine yayılacaksa, ki bunu yapması gerektiğini düşündü, o zaman bunu dijital biçimde yapması gerektiğini savundu. Başlıca endişeleri, AI paranoyasının dijital ütopyayı geciktireceği ve / veya Google'ın "kötü olmayın" sloganına ters düşecek şekilde AI'nın askeri olarak ele geçirilmesine neden olacağıydı. Elon geri adım atmaya devam etti ve Larry'den argümanlarının ayrıntılarını açıklığa kavuşturmasını istedi, örneğin neden dijital yaşamın önemsediğimiz her şeyi yok etmeyeceğine bu kadar güveniyordu. Larry zaman zaman Elon'u "türcü" olmakla suçladı: belirli yaşam formlarını sırf karbon bazlı olmaktansa silikon bazlı oldukları için aşağılık olarak görmek. 4. bölümden başlayarak bu ilginç konuları ve argümanları ayrıntılı olarak araştırmaya geri döneceğiz.

Larry, havuz kenarındaki o sıcak yaz gecesinde sayıca üstün görünse de, çok anlamlı bir şekilde savunduğu dijital ütopyacılığın birçok önde gelen destekçisi var. Robotik ve fütürist Hans Moravec, 1988 tarihli klasik kitabıyla tüm dijital ütopyacılara ilham verdi. *Akıl Çocukları,* mucit Ray Kurzweil tarafından sürdürülen ve rafine edilen bir gelenek. Pekiştirmeli öğrenme olarak bilinen AI alt alanının öncülerinden Richard Sutton, Porto Riko konferansımızda sizlere kısaca anlatacağım dijital ütopyacılığın tutkulu bir savunmasını verdi.

Tekno-şüpheciler

Bir diğer önde gelen düşünür grubu da yapay zeka hakkında endişelenmiyor, ancak tamamen farklı bir nedenden ötürü: süper insan YGZ'yi inşa etmenin o kadar zor olduğunu düşünüyorlar ki, yüzlerce yıl olmayacak ve bu yüzden endişelenmeyi aptalca görüyorlar. şimdi. Ben bunu şu şekilde düşünüyorum *tekno-şüpheci* Andrew Ng tarafından açıkça ifade edilen pozisyon: "Katil robotların artmasından korkmak, Mars'taki aşırı nüfus hakkında endişelenmek gibidir." Andrew, Çin'in Google'ı Baidu'da baş bilim adamıydı ve son zamanlarda onunla Boston'da bir konferansta konuştuğumda bu tartışmayı tekrarladı. Ayrıca bana, AI riski hakkında endişelenmenin, AI'nın ilerlemesini yavaşlatabilecek potansiyel olarak zararlı bir dikkat dağıtıcı olduğunu hissettiğini söyledi. Roomba robot elektrikli süpürgesinin ve Baxter endüstriyel robotunun arkasındaki eski MIT profesörü Rodney Brooks gibi diğer tekno-şüpheciler tarafından da benzer düşünceler dile getirildi. İlginç buluyorum ki, dijital ütopyacılar ve tekno-şüpheciler AI hakkında endişelenmememiz gerektiği konusunda hemfikir olsalar da, başka pek az şey üzerinde anlaşıyorlar. Ütopyacıların çoğu, insan seviyesinde YÜT'nin önümüzdeki yirmi ila yüz yıl içinde gerçekleşebileceğini düşünüyor. Tekno-şüphecilerin, bilgisiz bir rüya gibi gördüklerini, genellikle kehanet edilen tekilliği "ineklerin coşkusu" olarak alaya aldıkları. Rodney Brooks ile Aralık 2014'te bir doğum günü partisinde tanıştığımda bana bunun hayatım boyunca olmayacağından% 100 emin olduğunu söyledi. "% 99'u kastetmediğinden emin misin?" Diye yanıtladığı bir takip e-postasında sordum, "% 99 pürüz yok. 100%. Sadece olmayacak. "

Faydalı Yapay Zeka Hareketi

Stuart Russell ile Haziran 2014'te Paris'teki bir kafede ilk tanıştığımda, bana mükemmel bir İngiliz beyefendi olarak vurdu. Açık sözlü, düşünceli ve yumuşak dilli, ama gözlerinde maceracı bir parıltıyla, bana Jules Verne'in klasik 1873 romanındaki çocukluk kahramanım Phileas Fogg'un modern bir enkarnasyonu gibi geldi. *80 Günde Dünya Turu.* Yaşayan en ünlü yapay zeka araştırmacılarından biri olmasına rağmen, konuyla ilgili standart ders kitabının ortak yazarı olmasına rağmen, alçakgönüllülüğü ve sıcaklığı beni çok geçmeden rahatlattı. Bana yapay zekadaki ilerlemenin onu bu yüzyılda insan seviyesinde YÜZ'nin gerçek bir olasılık olduğuna ve umutlu olmasına rağmen iyi bir sonucun garanti edilmediğine nasıl ikna ettiğini anlattı. İlk önce cevaplamamız gereken çok önemli sorular vardı ve onlar o kadar zordu ki, şimdi onları araştırmaya başlamalıyız, böylece cevapları ihtiyaç duyduğumuzda hazır hale getirebilirdik.

Bugün, Stuart'ın görüşleri oldukça yaygın ve dünyadaki birçok grup, onun savunduğu yapay zeka güvenliği araştırmasının peşinde koşuyor. Ancak bu her zaman böyle değildi. İçinde bir makale *Washington post* 2015'i, AI güvenliği araştırmasının ana akım haline geldiği yıl olarak anıyor. Bundan önce, yapay zeka risklerine ilişkin konuşma, ana akım yapay zeka araştırmacıları tarafından genellikle yanlış anlaşıldı ve yapay zekanın ilerlemesini engellemeyi amaçlayan Luddite korkutucu çığırtkanlığı olarak reddedildi. Beşinci bölümde inceleyeceğimiz gibi, Stuart'ınkine benzer endişeler ilk olarak yarım asır önce, II.Dünya Savaşı sırasında Alman kodlarını çözmek için Turing ile birlikte çalışan bilgisayar öncüsü Alan Turing ve matematikçi Irving J. Good tarafından dile getirildi. Geçtiğimiz on yılda, bu tür konulardaki araştırmalar esas olarak profesyonel yapay zeka araştırmacısı olmayan bir avuç bağımsız düşünür tarafından gerçekleştirildi, örneğin Eliezer Yudkowsky, Michael Vassar ve Nick Bostrom. Çalışmalarının çoğu ana akım AI araştırmacısı üzerinde çok az etkisi oldu. Başarının uzun vadeli sonuçlarını düşünmek yerine, AI sistemlerini daha akıllı hale getirme konusundaki günlük görevlerine odaklanma eğilimindeydiler. Kimin endişe duyduğunu tanıdığım yapay zeka araştırmacılarının çoğu, alarmcı teknofoblar olarak algılanma korkusuyla bunu dile getirmekte tereddüt etti.

Bu kutuplaşmış durumun değişmesi gerektiğini hissettim, böylece tüm yapay zeka topluluğu yararlı yapay zekanın nasıl oluşturulacağıyla ilgili sohbete katılabilir ve onu etkileyebilirdi. Neyse ki yalnız değildim. 2014 baharında Future of Life Institute (FLI;

http://futureoflife.org) eşim Meia, fizikçi arkadaşım Anthony Aguirre, Harvard mezunu Viktoriya Krakovna ve Skype kurucusu Jaan Tallinn ile birlikte. Amacımız basitti: Hayatın geleceğinin var olmasını ve olabildiğince harika olmasını sağlamaya yardımcı olmak. Spesifik olarak, teknolojinin hayata ya daha önce hiç olmadığı gibi gelişme ya da kendi kendini yok etme gücü verdiğini hissettik ve ilkini tercih ettik.

İlk toplantımız 15 Mart'ta evimizde bir beyin fırtınası oturumuydu.
Boston bölgesinden yaklaşık otuz öğrenci, profesör ve diğer düşünürlerle 2014. Biyoteknoloji, nükleer silahlar ve iklim değişikliğine dikkat etmemiz gerekmesine rağmen, ilk ana hedefimizin AI-güvenlik araştırmasını ana akım haline getirmeye yardımcı olmak olması gerektiği konusunda geniş bir fikir birliği vardı. Kuarkların nasıl çalıştığını anlamaya yardımcı olduğu için Nobel Ödülü kazanan MIT fizik meslektaşım Frank Wilczek, konuya dikkat çekmek ve görmezden gelmeyi zorlaştırmak için bir köşe yazısı yazarak başlamamızı önerdi. Stuart Russell'a (henüz tanışmadığım) ve fizik meslektaşım Stephen Hawking'e ulaştım, ikisi de bana ve Frank'e yardımcı yazar olarak katılmayı kabul etti. Birçok düzenleme daha sonra, görüşümüz tarafından reddedildi *New York Times* ve diğer birçok ABD gazetesi, bu yüzden onu benim *Huffington Post* blog hesabı. Arianna Huffington kendisi e-posta ile göndermiş ve “bunu elde ettiği için çok heyecanlıyım! 1 numarada yayınlayacağız! ”Ve ön sayfanın üst kısmındaki bu yerleşim, Elon Musk, Bill Gates ve diğer teknoloji liderleri ile yılın geri kalanında süren yapay zeka güvenliği medyasında bir haber dalgasını tetikledi. Nick Bostrom'un kitabı *Süper zeka* o sonbaharda ortaya çıktı ve büyüyen kamusal tartışmayı daha da alevlendirdi.

FLI yararlı AI kampanyamızın bir sonraki amacı, dünyanın önde gelen AI araştırmacılarını yanlış anlamaların giderilebileceği, fikir birliğinin uydurulabileceği ve yapıcı planların yapılabileceği bir konferansa getirmekti. Böylesine şanlı bir kalabalığı, özellikle tartışmalı konu göz önüne alındığında, tanımadıkları yabancılar tarafından düzenlenen bir konferansa gelmeye ikna etmenin zor olacağını biliyorduk, bu yüzden elimizden geldiğince çabaladık: medyanın katılmasını yasakladık, yerini tespit ettik. Ocak ayında bir sahil beldesinde (Porto Riko'da), ücretsiz yaptık (Jaan Tallinn'in cömertliği sayesinde) ve bulabildiğimiz en endişe verici olmayan başlığı verdik: "Yapay Zekanın Geleceği: Fırsatlar ve Zorluklar. " En önemlisi, Stuart Russell ile birlikte çalıştık. Organizasyon komitesini hem akademi hem de sektörden bir grup AI liderini içerecek şekilde büyütebildiğimiz için teşekkür ederiz. Bunların arasında, AI'nın Go oyununda bile insanları yenebileceğini gösteren Google'ın DeepMind'ından Demis Hassabis de var. Demis'i daha çok tanıdıkça, sahip olduğunu daha çok anladım.

Yapay zekayı sadece güçlü kılmak değil, aynı zamanda onu faydalı kılmak için de hırs.

Sonuç, dikkate değer bir zihin buluşmasıydı (şekil 1.3). AI araştırmacıları en iyi ekonomistler, hukuk bilim adamları, teknoloji liderleri (Elon Musk dahil) ve diğer düşünürler (4. bölümün odak noktası olan “tekillik” terimini icat eden Vernor Vinge dahil) katıldı. Sonuç, en iyimser beklentilerimizi bile aştı. Belki güneş ışığı ve şarabın bir kombinasyonuydu ya da belki de tam zamanıydı: tartışmalı olmasına rağmen

açık bir mektupta kodladığımız dikkate değer bir fikir birliği ortaya çıktı. [2] Bu, gerçek bir AI'da kim kim olan dahil sekiz binden fazla kişi tarafından imzalanmasıyla sonuçlandı. Mektubun özü, yapay zekanın amacının yeniden tanımlanması gerektiğiydi: amaç, yönlendirilmemiş zeka değil, faydalı zeka yaratmak olmalıdır. Mektupta ayrıca, konferans katılımcılarının bu hedefi daha da ileriye götürmek konusunda hemfikir oldukları ayrıntılı bir araştırma konuları listesinden bahsedildi. Yararlı AI hareketi ana akım olmaya başlamıştı. Bundan sonraki ilerlemesini kitabın ilerleyen bölümlerinde izleyeceğiz.



Şekil 1.3: Ocak 2015 Porto Riko konferansı, AI ve ilgili alanlarda dikkate değer bir grup araştırmacıyı bir araya getirdi. Arka sıra, soldan sağa: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Diğer ayakta, soldan sağa: Marilyn Thompson, Rich Sutton, Alex Wissner- Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, Max Tegmark, Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel , James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Oturanlar: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. Kameranın arkasında: Anthony Aguirre (ve yanında oturan insan seviyesindeki zekanın da photoshop'unu yaptı). Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Oturanlar: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. Kameranın arkasında: Anthony Aguirre (ve yanında oturan insan seviyesindeki zekanın da photoshop'unu yaptı). Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Oturanlar: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. Kameranın arkasında: Anthony Aguirre (ve yanında oturan insan seviyesindeki zekanın da photoshop'unu yaptı). Viktoriya Krakovna, Meia

Konferanstan bir diğer önemli ders şuydu: YZ'nin başarısının ortaya çıkardığı sorular sadece entelektüel açıdan büyüleyici değil; aynı zamanda ahlaki açıdan da önemlidir, çünkü seçimlerimiz potansiyel olarak yaşamın tüm geleceğini etkileyebilir. İnsanlığın geçmiş seçimlerinin ahlaki önemi bazen harikaydı, ancak her zaman sınırlıydı: en büyük belalardan bile kurtulduk ve hatta en büyük imparatorluklar sonunda yıkıldı. Geçmiş nesiller, Güneş'in yarın doğacağı gibi, yarının insanlarının da sonsuz belalarla mücadele edeceğini biliyordu.

yoksulluk, hastalık ve savaş gibi. Ancak Porto Riko konuşmacılarından bazıları bu seferin farklı olabileceğini savundu: ilk defa, bu belaları kalıcı olarak sona erdirecek veya insanlığın kendisini sona erdirecek kadar güçlü bir teknoloji geliştirebileceğimizi söylediler. Dünyada ve belki de ötesinde daha önce hiç olmadığı kadar gelişen toplumlar ya da asla devrilmeyecek kadar güçlü Kafkavari bir küresel gözetim devleti yaratabiliriz.



Şekil 1.4: Medya, Elon Musk'u AI topluluğuyla sık sık anlaşmazlık halinde olarak tasvir etse de, aslında AI güvenliği araştırmasının gerekli olduğu konusunda geniş bir fikir birliği var. Ocak'ta burada
4 Ekim 2015, Yapay Zekayı Geliştirme Derneği başkanı Tom Dietterich, Elon'un çok daha önce finanse etmeyi taahhüt ettiği yeni yapay zeka güvenliği araştırma programı hakkındaki heyecanını paylaşıyor. FLI kurucuları Meia Chita-Tegmark ve Viktoriya Krakovna onların arkasında pusuda bekliyor.

Yanılgılar

Porto Riko'dan ayrıldığımda, orada yapay zekanın geleceği hakkında yaptığımız konuşmanın devam etmesi gerektiğine o kadar inandım çünkü en önemlisi bu zamanımızın konuşması. * 2 Hepimizin kolektif geleceği hakkında bir konuşma, bu yüzden yapay zeka araştırmacılarıyla sınırlı kalmamalı. Bu yüzden bu kitabı yazdım: Sevgili okurumun bu sohbete katılmanız umuduyla yazdım. Ne tür bir gelecek istiyorsun? Ölümcül otonom silahlar geliştirmeli miyiz? İş otomasyonu ile ne olmasını istersiniz? Bugünün çocuklarına hangi kariyer tavsiyelerini verirdiniz? Eski işlerin yerine yeni işleri mi yoksa herkesin boş vakit ve makine tarafından üretilen zenginlik dolu bir hayattan zevk aldığı işsiz bir toplumu mu tercih ediyorsunuz? Daha ileride, Life 3.0'ı yaratmamızı ve onu kozmosumuza yaymamızı ister misiniz? Akıllı makineleri kontrol edecek miyiz yoksa bizi kontrol edecekler mi? Akıllı makineler bizi değiştirecek mi, bizimle birlikte mi var olacak yoksa bizimle mi birleşecek? Yapay zeka çağında insan olmak ne anlama gelecek? Ne demek istedin

Bu kitabın amacı, bu sohbete katılmanıza yardımcı olmaktır. Bahsettiğim gibi, dünyanın önde gelen uzmanlarının aynı fikirde olmadığı büyüleyici tartışmalar var. Ama aynı zamanda insanların yanlış anladıkları ve birbirlerinin yanından geçtikleri pek çok sıkıcı sözde tartışma örneği de gördüm. Yanlış anlaşılmalara değil, ilginç tartışmalara ve açık sorulara odaklanmamıza yardımcı olmak için, en yaygın yanlış anlamalardan bazılarını açıklığa kavuşturarak başlayalım.

"Yaşam", "zeka" ve "bilinç" gibi terimler için yaygın olarak kullanılan birçok rakip tanım vardır ve birçok yanlış anlama, bir kelimeyi iki farklı şekilde kullandıklarının farkında olmayan insanlardan gelir. Senin ve benim bu tuzağa düşmediğimizden emin olmak için, bir kopya kağıdı koydum tablo 1.1 bu kitaptaki anahtar terimleri nasıl kullandığımı gösteriyor. Bu tanımlardan bazıları sadece daha sonraki bölümlerde uygun şekilde tanıtılacak ve açıklanacaktır. Lütfen tanımlarımın başkalarınınkinden daha iyi olduğunu iddia etmediğime dikkat edin - sadece ne demek istediğimi netleştirerek kafa karışıklığını önlemek istiyorum. Genelde insan merkezli önyargıdan kaçınan ve makinelere olduğu kadar insanlara da uygulanabilen geniş tanımları tercih ettiğimi göreceksiniz. Lütfen şimdi kopya kağıdını okuyun ve daha sonra geri dönün ve eğer sözlerinden birini nasıl kullandığıma şaşırırsanız, özellikle 4-8. Bölümlerde kontrol edin.

Terminoloji Hile Sayfası	
Hayat	Karmaşıklığını koruyabilen ve çoğaltabilen süreç
Hayat 1.0	Donanım ve yazılımını geliştiren yaşam (biyolojik aşama)
Yaşam 2.0	Donanımını geliştiren ancak yazılımlarının çoğunu tasarlayan yaşam (kültürel aşama)
Yaşam 3.0	Donanım ve yazılımını tasarlayan yaşam (teknolojik aşama)
Zeka	Karmaşık hedeflere ulaşma yeteneği
Yapay Zeka (AI)	Biyolojik olmayan zeka
Dar zeka	Dar bir hedef kümesine ulaşma becerisi, ör. satranç oynamak veya araba kullanmak
Genel zeka	Öğrenme dahil hemen hemen her hedefi gerçekleştirme yeteneği
Evrensel zeka	Verilere ve kaynaklara erişim verilen genel istihbarat elde etme yeteneği
[İnsan düzeyinde] Yapay Genel İstihbarat (AGI)	En az insanlar kadar herhangi bir bilişsel görevi yerine getirme yeteneği
İnsan düzeyinde AI	AGI
Güçlü AI	AGI
Süper zeka	İnsan seviyesinin çok ötesinde genel zeka
Medeniyet	Etkileşen akıllı yaşam formları grubu Öznel
Bilinç	deneyim
Qualia	Bireysel öznel deneyim örnekleri
Etik	Nasıl davranmamız gerektiğini yöneten ilkeler

Teleoloji	Şeylerin sebeplerinden ziyade amaçları veya amaçları açısından açıklaması
Hedef odaklı davranış	Davranış, nedeninden ziyade etkisiyle daha kolay açıklanır
Bir hedefe sahip olmak	Hedefe yönelik davranış sergilemek
Amaç sahibi olmak	Kendi veya başka bir varlığın hedeflerine hizmet etmek
Dostu AI	Hedefleri bizimkilerle uyumlu olan süper zeka
Yarı robot	İnsan-makine karması
Zeka patlama	Hızlı bir şekilde süper zekaya yol açan özyinelemeli kendini geliştirme
Tekillik	İstihbarat patlaması
Evren	Büyük Patlamamızdan bu yana 13,8 milyar yıl boyunca ışığın bize ulaşmak için zamanının olduğu uzay bölgesi

Tablo 1.1: YZ ile ilgili birçok yanlış anlama, yukarıdaki kelimeleri farklı şeyler ifade etmek için kullanan kişilerden kaynaklanmaktadır. İşte onları bu kitapta kastettiğim şey. (Bu tanımlardan bazıları yalnızca uygun şekilde tanıtılacak ve sonraki bölümlerde açıklanacaktır.)

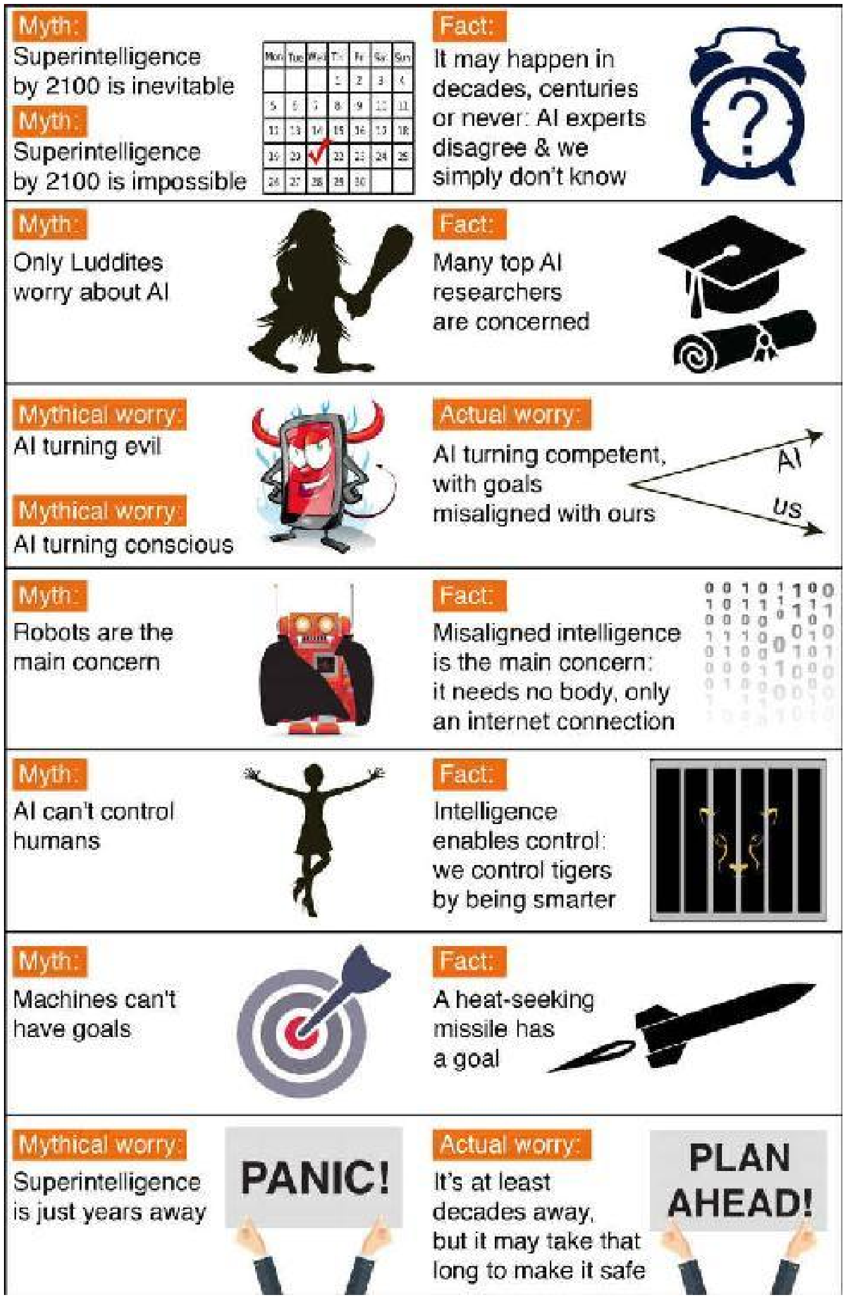
Terminoloji konusundaki kafa karışıklığına ek olarak, birçok AI konuşmasının basit yanlış anlamalar nedeniyle raydan çıktığını da gördüm. En yaygın olanları açıklayalım.

Zaman Çizelgesi Efsaneleri

İlki, zaman çizelgesine şekil 1.2 : Makinelerin insan seviyesinde AGI'nin yerini alması ne kadar sürer? Burada yaygın bir yanılgı, yanıtı büyük bir kesinlikle bildiğimizdir.

Popüler bir efsane, bu yüzyılda insanüstü AGI alacağımızı bilmemizdir. Aslında, tarih teknolojik abartılarla doludur. Şimdiye kadar sahip olacağımıza söz verilen füzyon santralleri ve uçan arabalar nerede? AI da geçmişte, alanın bazı kurucuları tarafından bile defalarca abartıldı: örneğin, John McCarthy ("yapay zeka" terimini bulan), Marvin Minsky, Nathaniel Rochester ve Claude Shannon bunu fazlasıyla iyimser olarak yazdı. Taş devri bilgisayarlarıyla iki ayda neler başarılabileceğine dair tahmin: “1956 yazında Dartmouth College'da 2 aylık 10 kişilik bir yapay zeka çalışmasının yapılmasını öneriyoruz… Nasıl yapılacağını bulmak için bir girişimde bulunulacak. makinelerin dili kullanmasını, soyutlamalar ve kavramlar oluşturmasını, artık insanlar için ayrılan problem türlerini çözmesini ve kendilerini geliştirmesini sağlayın.

Öte yandan, popüler bir karşı-efsane, *alışkanlık* bu yüzyılda insanüstü AGI olsun. Araştırmacılar, insanüstü YGZ'den ne kadar uzak olduğumuza dair geniş bir tahmin yelpazesi yaptılar, ancak bu tür tekno-şüpheci tahminlerin kasvetli geçmişi göz önüne alındığında, bu yüzyılda olasılığın sıfır olduğunu kesinlikle büyük bir güvenle söyleyemeyiz. Örneğin, zamanının tartışmasız en büyük nükleer fizikçisi Ernest Rutherford, 1933'te - Leo Szilard'ın nükleer zincir reaksiyonunu icat etmesinden yirmi dört saatten kısa bir süre önce - nükleer enerjinin "kaçak içki" olduğunu ve 1956'da Astronom Royal Richard Woolley uzay yolculuğu hakkında konuşmaya "mutlak sintine" denirdi. Bu efsanenin en uç biçimi, insanüstü AGI'nin fiziksel olarak imkansız olduğu için asla gelmeyeceğidir. Bununla birlikte, fizikçiler, bir beynin güçlü bir bilgisayar görevi görecek şekilde düzenlenmiş kuark ve elektronlardan oluştuğunu bilirler.



Şekil 1.5: Süper zeki AI ile ilgili yaygın efsaneler.

Yapay zeka araştırmacılarına kaç yıl sonra en az% 50 olasılıkla insan düzeyinde AGI'ye sahip olacağımızı düşündüklerini soran bir dizi anket yapıldı ve tüm bu anketler aynı sonuca sahip:

katılmıyorum, bu yüzden bilmiyoruz. Örneğin, Porto Riko AI konferansındaki AI araştırmacılarının böyle bir anketinde, ortalama (medyan) cevap 2055 yılına kadardı, ancak bazı araştırmacılar yüzlerce yıl veya daha fazlasını tahmin ettiler.

Yapay zeka konusunda endişelenen insanların, bunun sadece birkaç yıl uzakta olduğunu düşündükleri ile ilgili bir efsane de var. Aslında, insanüstü AGI hakkında kaygılanan kayıtlardaki çoğu insan, sanırım hala en az on yıl uzakta. Ama% 100 olmadığımız sürece *Elbette* Bu yüzyılda olmayacağını, olasılığa hazırlanmak için şimdi güvenlik araştırmasına başlamak akıllıca olacaktır. Bu kitapta göreceğimiz gibi, güvenlik sorunlarının birçoğu o kadar zor ki çözmeleri onlarca yıl sürebilir, bu nedenle Red Bull içen bazı programcıların insan seviyesinde geçiş yapmaya karar vermesinden önceki gece yerine bunları araştırmaya hemen başlamak akıllıca olacaktır. AGI.

Tartışma Efsaneleri

Bir başka yaygın yanlış anlama da, YZ hakkında endişeleri olan ve YZ güvenliği araştırmasını savunanların yalnızca YZ hakkında fazla bir şey bilmeyen Ludditler olduğudur. Stuart Russell Porto Riko konuşması sırasında bundan bahsettiğinde, seyirci yüksek sesle güldü. Bununla ilgili bir yanlış anlama, AI güvenliği araştırmasını desteklemenin büyük ölçüde tartışmalı olduğudur. Aslında, yapay zeka güvenlik araştırmalarına yapılan mütevazı bir yatırımı desteklemek için, insanların risklerin yüksek olduğuna, sadece ihmal edilemez olduğuna ikna edilmesine gerek yoktur, tıpkı ev sigortasına yapılan mütevazı bir yatırımın, göz ardı edilemez bir olasılıkla haklı gösterilmesi gibi, ev yanıyor.

Kişisel analizim, medyanın yapay zeka güvenliği tartışmasını gerçekte olduğundan daha tartışmalı hale getirdiği yönünde. Sonuçta, korku satıyor ve yaklaşan kıyamet ilan etmek için bağlam dışı alıntılar kullanan makaleler, incelikli ve dengeli olanlardan daha fazla tıklama üretebilir. Sonuç olarak, birbirlerinin konumlarını yalnızca medyadan alıntılardan bilen iki kişi, gerçekte yaptıklarından daha fazla aynı fikirde olmadıklarını düşüneceklerdir. Örneğin, Bill Gates'in konumu hakkında tek bilgisi İngiliz gazetelerinden gelen bir tekno-şüpheci, yanlışlıkla süper zekanın yakın olacağına inandığını düşünebilir. Benzer şekilde, faydalı AI hareketinde, Andrew Ng'nin konumu hakkında yukarıda bahsedilen Mars'taki aşırı nüfusla ilgili alıntı dışında hiçbir şey bilmeyen biri yanlışlıkla AI güvenliğini umursamadığını düşünebilir. Aslında,

Risklerin Neler Olduğuna Dair Mitler

Bu manşeti görünce gözlerimi devirdim *Günlük posta:* [3] "Stephen Hawking, Robotların Yükselişinin İnsanlık İçin Felaket Olabileceği Uyardı." Kaç tane benzer makale gördüğümün sayısını kaybettim. Tipik olarak, onlara silah taşıyan kötü görünümlü bir robot eşlik ediyor ve robotların yükselip bizi öldürmeleri konusunda endişelenmemizi öneriyorlar çünkü bunlar bilinçli ve / veya kötü oluyor. Daha hafif bir not olarak, bu tür makaleler aslında oldukça etkileyicidir, çünkü yapay zeka meslektaşlarımın senaryosunu kısa ve öz bir şekilde özetlerler. *yapma* endişelenmek. Bu senaryo, üç farklı yanılgıyı birleştirir:

bilinç, kötülük ve *robotlar* sırasıyla.

Yolda sürerseniz, öznel bir renk, ses deneyimi yaşarsınız. Peki sürücüsüz bir arabanın öznel bir deneyimi var mı? Kendi kendine giden bir araba olmak herhangi bir şeymiş gibi hissettiriyor mu, yoksa herhangi bir öznel deneyimi olmayan bilinçsiz bir zombi gibi mi? Bu bilinç gizemi kendi başına ilginç olsa da ve 8. bölümü ona ayıracak olsak da, AI riskiyle ilgisi yok. Sürücüsüz bir arabaya çarparsanız, onun öznel olarak bilinçli hissetmesi sizin için hiçbir fark yaratmaz. Aynı şekilde, biz insanları etkileyecek olan şey, süper zeki yapay zeka *yapar* öznel olarak nasıl hissettirdiği değil.

Makinelerin kötülüğe dönüşmesi korkusu bir başka kırmızı ringa balığı. Asıl endişe kötü niyet değil, yeterliliktir. Süper zeki bir YZ, ne olursa olsun, hedeflerine ulaşmada tanımı gereği çok iyidir, bu nedenle hedeflerinin bizimkilerle uyumlu olmasını sağlamamız gerekir. Muhtemelen kötülükten karıncalara basan bir karınca düşmanı değilsiniz, ancak bir hidroelektrik yeşil enerji projesinden sorumluysanız ve bölgede sular altında kalacak bir karınca yuvası varsa, karıncalar için çok kötü. Yararlı AI hareketi, insanlığı bu karıncaların yerine yerleştirmekten kaçınmak istiyor.

Bilinç yanılgısı, makinelerin hedefleri olamayacağı mitiyle ilgilidir. Makinelerin, hedefe yönelik davranış sergileme gibi dar anlamda hedefleri olduğu açıktır: Isı arayan bir füzenin davranışı, en ekonomik olarak bir hedefi vurma hedefi olarak açıklanır. Hedefleri sizinkiyle yanlış hizalanmış bir makine tarafından tehdit edildiğinizi hissediyorsanız, makinenin bilinçli olup olmadığı ve bir amaç duygusu yaşayıp yaşamadığı değil, tam da bu dar anlamdaki hedefleri sizi rahatsız eder. Eğer o ısı arayan füze sizi kovalıyor olsaydı, muhtemelen "Endişelenmiyorum, çünkü makinelerin hedefleri olamaz!" Diye bağırmazdınız.

Rodney Brooks ve diğer robotik öncülerine sempati duyuyorum, çünkü bazı gazeteciler takıntılı bir şekilde robotlara odaklanmış görünüyor ve makalelerinin çoğunu parlak kırmızı gözlü kötü görünümlü metal canavarlarla süslüyor. Aslında, faydalı AI hareketinin ana endişesi robotlarla değil, zekanın kendisiyle ilgilidir: özellikle hedefleri bizimkilerle yanlış hizalanmış zeka. Bizim sorunumuza neden olmak için, bu tür yanlış hizalanmış istihbaratın robotik bir vücuda ihtiyacı yoktur, sadece bir internet bağlantısı vardır - 4. bölümde bunun finansal piyasaları alt etmeyi, insan araştırmacıları alt etmeyi, insan liderleri alt etmeyi ve hatta yapamayacağımız silahları geliştirmeyi nasıl sağlayabileceğini inceleyeceğiz. anlama. Robotlar inşa etmek fiziksel olarak imkansız olsa bile, *Neuromancer.*

Robot yanılgısı, makinelerin insanları kontrol edemeyeceği efsanesiyle ilgilidir. Zeka kontrolü sağlar: İnsanlar kaplanları daha güçlü olduğumuz için değil, daha zeki olduğumuz için kontrol eder. Bu, gezegenimizdeki en akıllı konumumuzu terk edersek, kontrolü de bırakmamızın mümkün olduğu anlamına gelir.

Şekil 1.5 tüm bu yaygın yanlış anlamaları özetliyor, böylece onlardan bir kez ve tamamen vazgeçebilir ve arkadaşlarımızla ve meslektaşlarımızla yaptığımız tartışmaları birçok meşru tartışmaya odaklayabiliriz - ki göreceğimiz gibi, hiçbir eksiklik yok!

Öndeki yol

Bu kitabın geri kalanında, sen ve ben yaşamın geleceğini AI ile birlikte keşfedeceğiz. Bu zengin ve çok yönlü konuyu, önce kavramsal ve kronolojik olarak hayatın tüm öyküsünü kavramsal ve kronolojik olarak keşfederek, ardından hedefleri, anlamı ve istediğimiz geleceği yaratmak için hangi eylemleri gerçekleştireceğimizi keşfederek organize bir şekilde gezelim.

2. bölümde, zekanın temellerini ve görünüşte aptal olan maddenin hatırlamak, hesaplamak ve öğrenmek için nasıl yeniden düzenlenebileceğini keşfedeceğiz. Geleceğe doğru ilerlerken, hikayemiz belirli kilit soruların cevaplarıyla tanımlanan birçok senaryoya ayrılıyor. Şekil 1.6 Potansiyel olarak her zamankinden daha gelişmiş yapay zekaya doğru zamanda ilerledikçe karşılaşacağımız temel soruları özetliyor.

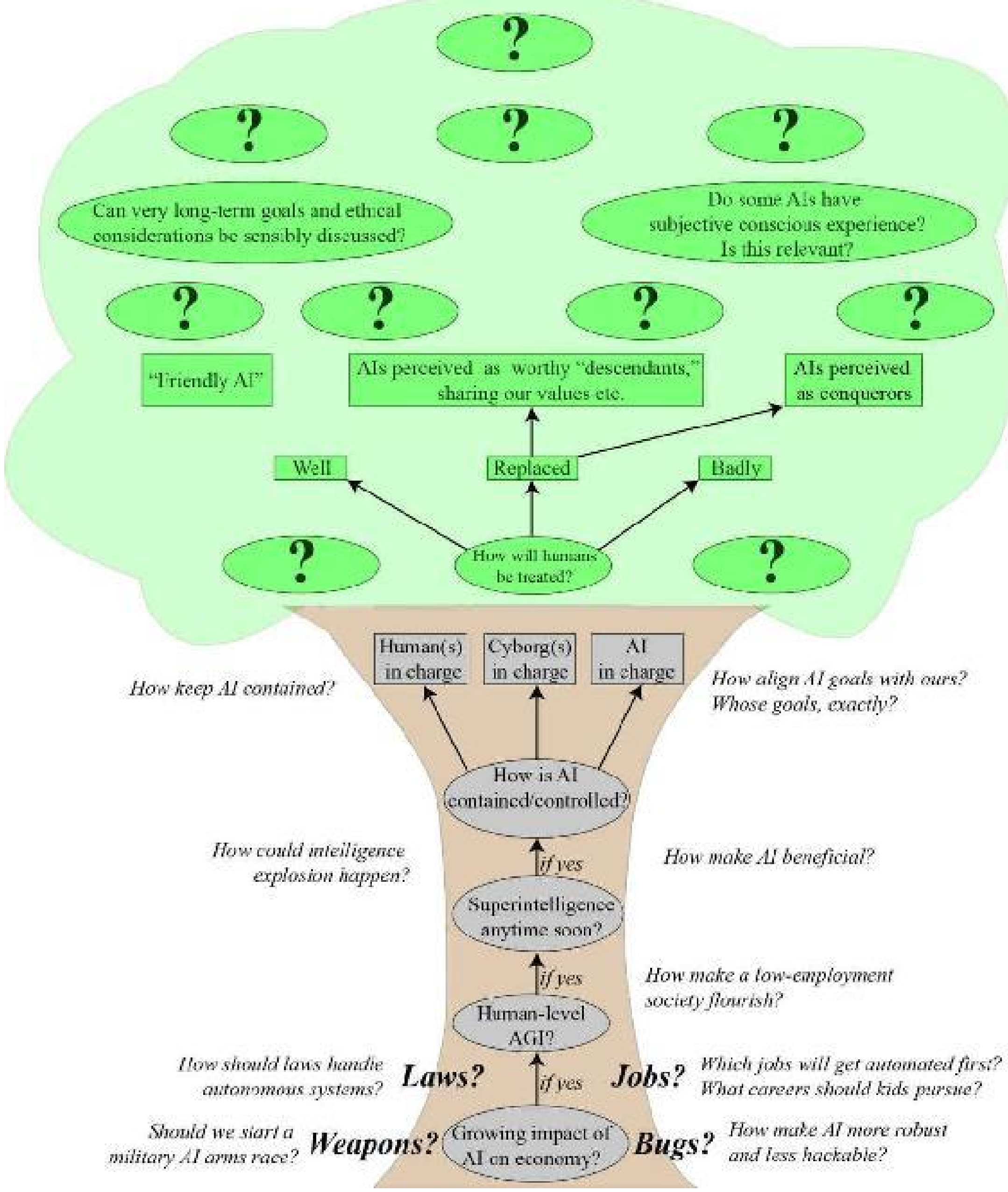
Şu anda, bir yapay zeka silahlanma yarışına başlayıp başlamama seçimi ve yarının yapay zeka sistemlerinin hatasız ve sağlam nasıl yapılacağına dair sorularla karşı karşıyayız. Yapay zekanın ekonomik etkisi artmaya devam ederse, yasalarımızı nasıl modernize edeceğimize ve çocuklara yakında otomatik hale gelecek işlerden kaçınabilmeleri için hangi kariyer tavsiyelerini vereceğimize de karar vermeliyiz. Bu tür kısa vadeli soruları 3. bölümde inceleyeceğiz.

Yapay zeka ilerlemesi insan seviyelerinde devam ederse, o zaman kendimize bunun yararlı olmasını nasıl sağlayacağımızı ve işsiz gelişen bir boş zaman toplumu yaratıp yaratamayacağımızı da sormalıyız. Bu aynı zamanda bir istihbarat patlamasının mı yoksa yavaş ama istikrarlı büyümenin AGI'yi insan seviyelerinin çok ötesine taşıyıp taşımayacağı sorusunu da gündeme getiriyor. Bölüm 4'te bu tür çok çeşitli senaryoları araştırıyoruz ve 5. bölümde tartışmalı distopikten tartışmalı ütopik olana kadar olasılıkların spektrumunu araştırıyoruz. Kim sorumlu - insanlar mı, yapay zeka mı yoksa cyborglar mı? İnsanlara iyi mi yoksa kötü mü davranılıyor? Değiştiriliyor muyuz ve eğer öyleyse, değiştirmelerimizi fatihler veya değerli torunlar olarak görüyor muyuz? Kişisel olarak 5. bölüm senaryolarından hangisini tercih ettiğinizi çok merak ediyorum! Bir web sitesi kurdum http://AgeOfAi.org , görüşlerinizi paylaşabileceğiniz ve sohbete katılabileceğiniz yer.

Son olarak, evrenimizdeki yaşamın nihai sınırları zeka tarafından değil, fizik yasaları tarafından belirlendiğinden, ironik bir şekilde, önceki bölümlerden daha güçlü sonuçlar çıkarabileceğimiz 6. bölümde milyarlarca yılı geleceğe atıyoruz.

Zeka tarihi araştırmamızı tamamladıktan sonra,

kitabın geri kalanı, hangi geleceği hedefleyeceğimizi ve oraya nasıl ulaşılacağını düşünmeye. Soğuk gerçekleri amaç ve anlam sorularıyla ilişkilendirebilmek için, 7. bölümdeki hedeflerin fiziksel temelini ve 8. bölümdeki bilinci keşfediyoruz. Son olarak, sonsözde, geleceği yaratmaya yardımcı olmak için şu anda neler yapılabileceğini araştırıyoruz. istemek.



Şekil 1.6: Hangi AI sorularının ilginç olduğu, AI'nın ne kadar gelişmiş olduğuna ve geleceğimizin hangi dalda olduğuna bağlıdır.

Atlamadan hoşlanan bir okuyucuysanız, çoğu bölüm, terminolojiyi ve tanımları bir kez sindirdikten sonra nispeten bağımsızdır.

ilk bölüm ve bir sonraki bölümün başlangıcı. Bir AI araştırmacısıysanız, isteğe bağlı olarak, ilk istihbarat tanımları haricinde bölüm 2'nin tümünü atlayabilirsiniz. Yapay zeka konusunda yeniyseniz, 2. ve 3. bölümler, neden 4'ten 6'ya kadar olan bölümlerin imkansız bilim kurgu olarak önemsiz bir şekilde göz ardı edilemeyeceğine dair argümanlar verecektir.

Şekil 1.7 çeşitli bölümlerin gerçeklerden spekülatife doğru yelpazenin neresinde olduğunu özetler.

		Short Chapter Title	*Topic*	*Status*
		Prelude: Tale of the Omega Team	Food for thought	Extremely Speculative
The history of intelligence	1	The Conversation	Key ideas, terminology	Not very speculative
	2	Matter Turns Intelligent	Fundamentals of intelligence	
	3	AI, Economics, Weapons & Law	Near future	
	4	Intelligence Explosion?	Superintelligence scenarios	Extremely Speculative
	5	Aftermath	Subsequent 10,000 years	
	6	Our Cosmic Endowment	Subsequent billions of years	
The history of meaning	7	Goals	History of goals-oriented behavior	Not very speculative
	8	Consciousness	Natural & artificial consciousness	Speculative
		Epilogue: Tale of the FLI Team	What should we do?	Not very speculative

Şekil 1.7: Kitabın Yapısı

Büyüleyici bir yolculuk bizi bekliyor. Hadi başlayalım!

ALT ÇİZGİ:

- Karmaşıklığını koruyabilen ve çoğaltabilen bir süreç olarak tanımlanan yaşam, üç aşamada gelişebilir: donanım ve yazılımın geliştiği biyolojik bir aşama (1.0), yazılımını tasarlayabildiği kültürel bir aşama (2.0) (öğrenme yoluyla) ve donanımını da tasarlayabileceği, kendi kaderinin efendisi haline gelebileceği teknolojik bir aşama (3.0).

- Yapay zeka, bu yüzyılda Life 3.0'ı başlatmamızı sağlayabilir ve hangi geleceği hedeflememiz gerektiği ve bunun nasıl başarılabileceği konusunda büyüleyici bir konuşma başladı. Tartışmada üç ana kamp var: tekno-şüpheciler, dijital ütopyacılar ve faydalı-AI hareketi.

- Tekno-şüpheciler, insanüstü YGZ'yi inşa etmeyi o kadar zor görüyor ki, yüzlerce yıl olmayacak, bu da bunun (ve Life 3.0'ın) hakkında endişelenmesini aptalca yapıyor.

- Dijital ütopyacılar, onu bu yüzyıl gibi görüyorlar ve Life 3.0'ı, kozmik evrimin doğal ve arzulanan bir sonraki adımı olarak gördükleri için tüm kalbiyle hoş karşıladılar.

- Faydalı AI hareketi de bu yüzyılın olası olduğunu düşünüyor, ancak iyi bir sonucun garanti edildiği gibi değil, AI güvenliği araştırması şeklinde sıkı çalışma ile sağlanması gereken bir şey olarak görüyor.

- Dünyanın önde gelen uzmanlarının aynı fikirde olmadığı bu tür meşru tartışmaların ötesinde, yanlış anlamaların neden olduğu sıkıcı sözde tartışmalar da vardır. Örneğin, sizin ve başkahramanın bu kelimeleri aynı anlama gelmek için kullandığından emin olmadan önce asla "yaşam", "zeka" veya "bilinç" hakkında tartışarak zaman kaybetmeyin! Bu kitap aşağıdaki tanımları kullanır tablo 1.1 .

- Ayrıca aşağıdaki konulardaki yaygın yanlış anlamalara dikkat edin: şekil 1.5 : "2100'e kadar süper zeka kaçınılmaz / imkansız." "Yapay zeka konusunda yalnızca Ludditler endişeleniyor." "Sorun, yapay zekanın kötülüğe ve / veya bilinçli hale getirilmesiyle ilgili ve sadece yıllar sonra." "Robotlar ana sorun." "Yapay zeka insanları kontrol edemez ve hedefleri olamaz."

- 2'den 6'ya kadar olan bölümlerde, zekanın hikayesini milyarlarca yıl önceki mütevazı başlangıcından, bugünden milyarlarca yıl sonraki olası kozmik geleceklere kadar keşfedeceğiz. Önce işler, yapay zeka silahları ve insan seviyesinde AGI arayışı gibi kısa vadeli zorlukları araştıracağız, ardından akıllı makineler ve / veya insanlarla olası geleceklerin büyüleyici bir yelpazesi için olasılıkları keşfedeceğiz. Hangi seçenekleri tercih edeceğinizi merak ediyorum!

- 7'den 9'a kadar olan bölümlerde, soğuk olgusal tanımlamalardan hedeflerin, bilincin ve anlamın araştırılmasına geçeceğiz ve istediğimiz geleceği yaratmaya yardımcı olmak için şu anda neler yapabileceğimizi araştıracağız.

- Yapay zeka ile yaşamın geleceği hakkındaki bu sohbeti zamanımızın en önemli biri olarak görüyorum - lütfen katılın!

* 1 Hayat neden daha karmaşık hale geldi? Evrim, çevresindeki düzenleri tahmin etmek ve kullanmak için yeterince karmaşık olan yaşamı ödüllendirir, böylece daha karmaşık bir ortamda, daha karmaşık ve zeki yaşam gelişecektir. Şimdi bu daha akıllı yaşam, rekabet halindeki yaşam formları için daha karmaşık bir ortam yaratıyor, bu da daha karmaşık hale gelecek ve sonunda son derece karmaşık bir yaşam ekosistemi yaratacak.

* 2 AI görüşmesi hem aciliyet hem de etki açısından önemlidir. Elli ila iki yüz yıl içinde hasara yol açabilecek iklim değişikliği ile karşılaştırıldığında, birçok uzman AI'nın on yıllar içinde daha büyük bir etkiye sahip olmasını ve potansiyel olarak bize iklim değişikliğini hafifletmek için teknoloji vermesini bekliyor. Savaşlar, terörizm, işsizlik, yoksulluk, göç ve sosyal adalet sorunları ile karşılaştırıldığında, yapay zekanın yükselişi daha büyük bir genel etkiye sahip olacaktır - aslında, bu kitapta tüm bu konularda olanlara nasıl daha iyi veya daha iyi hükmedebileceğini inceleyeceğiz. daha kötüsü için.

Bölüm 2

Madde Akıllı Oluyor

Hidrojen… yeterli zaman verildiğinde insana dönüşür.

Edward Robert Harrison, 1995

Büyük Patlamamızdan bu yana 13,8 milyar yıl boyunca yaşanan en çarpıcı gelişmelerden biri, aptal ve cansız maddenin zekaya dönüşmesidir. Bu nasıl olabilir ve gelecekte işler ne kadar akıllı hale gelebilir? Evrenimizdeki zekanın tarihi ve kaderi hakkında bilimin söylemesi gereken nedir? Bu soruların üstesinden gelmemize yardımcı olmak için bu bölümü zekanın temellerini ve temel yapı taşlarını keşfetmeye ayıralım. Bir damla maddenin zeki olduğunu söylemek ne anlama geliyor? Bir nesnenin hatırlayabildiğini, hesaplayabildiğini ve öğrenebileceğini söylemek ne demektir?

Zeka Nedir?

Karım ve ben yakın zamanda İsveç Nobel Vakfı tarafından düzenlenen yapay zeka konulu bir sempozyuma katılma şansına sahip olduk ve önde gelen YZ araştırmacılarından bir panelden zekayı tanımlamaları istendiğinde, uzunca bir fikir birliğine varmadan tartıştılar. Bunu oldukça komik bulduk: zeki zeka araştırmacıları arasında bile zekanın ne olduğu konusunda bir anlaşma yok! Dolayısıyla, zekanın tartışmasız "doğru" bir tanımı yoktur. Bunun yerine, mantık, anlama, planlama, duygusal bilgi, öz farkındalık, yaratıcılık, problem çözme ve öğrenme kapasitesi dahil olmak üzere birçok rakip var.

Zekanın geleceğini araştırırken, şimdiye kadar var olan zeka türleriyle sınırlı olmayan, azami ölçüde geniş ve kapsayıcı bir bakış açısı benimsemek istiyoruz. Bu yüzden son bölümde verdiğim tanım ve bu kitap boyunca bu sözcüğü kullanma şeklim çok geniştir:

zeka = *karmaşık hedeflere ulaşma yeteneği*

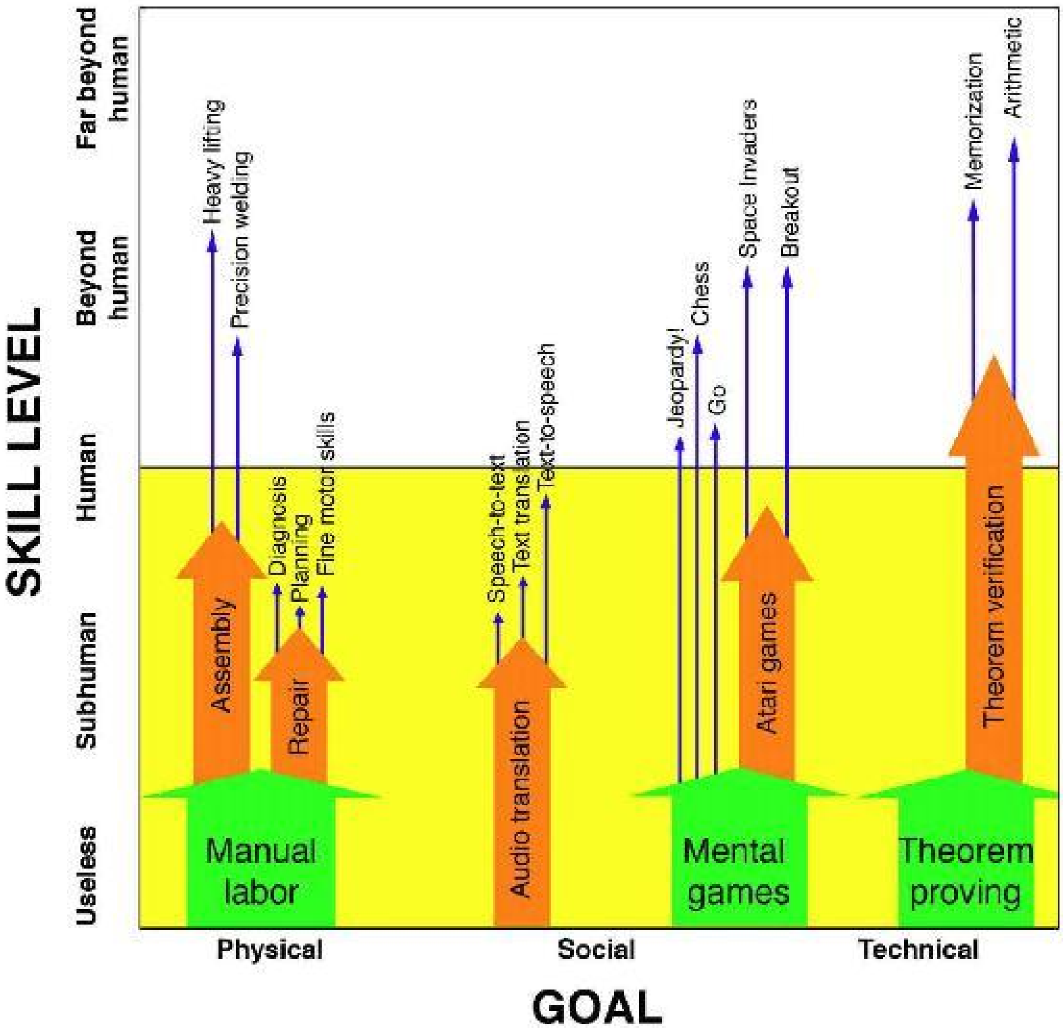
Anlama, öz farkındalık, problem çözme, öğrenme vb. Kişinin sahip olabileceği karmaşık hedeflere örnekler olduğu için bu, yukarıda belirtilen tüm tanımları içerecek kadar geniştir. Ayrıca, kapsayacak kadar geniştir.
Oxford Sözlüğü tanım - “bilgi ve becerileri edinme ve uygulama yeteneği” - bilgi ve becerileri uygulama hedefi olabileceğinden.

Pek çok olası hedef olduğundan, birçok olası zeka türü vardır. Tanımımıza göre, bu nedenle insanların, insan olmayan hayvanların veya makinelerin zekasını tek bir sayı ile ölçmenin bir anlamı yoktur.

IQ gibi. * [1] Hangisi daha akıllı: sadece satranç oynayabilen bir bilgisayar programı mı yoksa sadece Go oynayabilen bir bilgisayar programı mı? Doğrudan karşılaştırılamayacak farklı şeylerde iyi oldukları için buna mantıklı bir cevap yok. Bununla birlikte, üçüncü bir programın, başarmada en azından onlar kadar iyiyse, diğerlerinden daha akıllı olduğunu söyleyebiliriz. *herşey* hedefler ve kesinlikle daha iyi

en az bir (diyelim ki satrançta kazanmak).

Ayrıca, bir şeyin sınırda olan durumlarda akıllı olup olmadığı konusunda tartışmak pek mantıklı değildir, çünkü yetenek bir spektrumda gelir ve mutlaka ya hep ya hiç özelliği değildir. Konuşma amacına ulaşmak için hangi insanlar yetenekli? Yenidoğanlar? Hayır. Radyo sunucuları? Evet. Peki ya on kelime konuşabilen küçük çocuklar? Ya da beş yüz kelime? Çizgiyi nereye çekersiniz? Yukarıdaki tanımda kasıtlı olarak belirsiz olan "karmaşık" kelimesini kullandım, çünkü zeka ve zeka dışı arasında yapay bir çizgi çizmeye çalışmak çok ilginç değil ve farklı hedeflere ulaşma becerisinin derecesini basitçe ölçmek daha kullanışlı. .



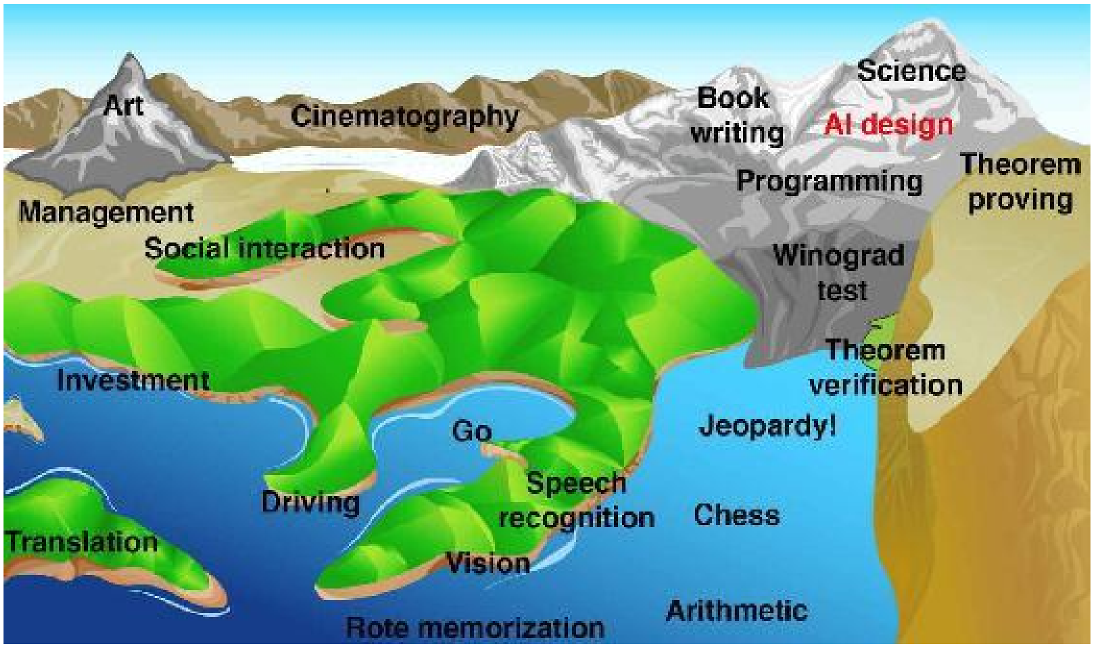
Şekil 2.1: Karmaşık hedefleri gerçekleştirme yeteneği olarak tanımlanan zeka, tek bir IQ ile ölçülemez, yalnızca tüm hedeflerdeki bir yetenek spektrumuyla ölçülebilir. Her ok, günümüzün en iyi yapay zeka sistemlerinin çeşitli hedeflere ulaşmada ne kadar yetenekli olduğunu gösteriyor ve günümüzün yapay zekasının *dar,* her sistem yalnızca çok özel hedeflere ulaşabilir. Buna karşılık, insan zekası oldukça geniştir: sağlıklı bir çocuk neredeyse her şeyde daha iyi olmayı öğrenebilir.

Farklı zekaları bir sınıflandırmada sınıflandırmak için bir başka önemli ayrım şudur: *dar* ve *kalın* zeka. 1997'de satranç şampiyonu Garry Kasparov'u tahttan indiren IBM'in Deep Blue satranç bilgisayarı, satranç oynamanın çok dar görevini başardı - etkileyici donanım ve yazılımına rağmen, dört yaşındaki birini tikte bile yenemedi. tac-toe.

Google DeepMind'ın DQN AI sistemi, biraz daha geniş bir hedef yelpazesine ulaşabilir: düzinelerce farklı eski Atari bilgisayar oyununu insan seviyesinde veya daha iyi oynayabilir. Buna karşılık, insan zekası şimdiye kadar benzersiz bir şekilde geniştir ve göz kamaştırıcı bir beceri yelpazesine hakim olabilir. Yeterli eğitim süresi verilen sağlıklı bir çocuk, yalnızca *hiç* oyun, aynı zamanda herhangi bir dilde, sporda veya meslekte. Günümüzde insanların ve makinelerin zekasını karşılaştırdığımızda, biz insanlar geniş bir alanda kazançlı çıkarken, makineler küçük ama artan sayıda dar alanda bizi geride bırakırken, şekil 2.1 . Yapay zeka araştırmasının kutsal kâsesi, "genel YZ" oluşturmaktır (daha iyi bilinen adıyla *yapay genel zeka,* AGI) maksimum derecede geniş: öğrenme dahil hemen hemen her hedefi gerçekleştirebilir. Bunu 4. bölümde ayrıntılı olarak inceleyeceğiz. "AGI" terimi, yapay zeka araştırmacıları Shane Legg, Mark Gubrud ve Ben Goertzel tarafından popüler hale getirildi. *insan seviyesi* yapay genel

zeka: en azından insanlar kadar herhangi bir hedefi gerçekleştirme yeteneği. [1] Tanımlarına sadık kalacağım, bu nedenle kısaltmayı açıkça nitelemediğim sürece (örneğin, "insanüstü AGI" yazarak), "AGI" yi "insan düzeyinde" kısaltma olarak kullanacağım
AGI. " * [2]

"Zeka" kelimesi olumlu çağrışımlara sahip olma eğiliminde olsa da, onu tamamen değerden bağımsız bir şekilde kullandığımızı belirtmek önemlidir: bu hedeflerin iyi veya kötü olarak kabul edilmesine bakılmaksızın karmaşık hedeflere ulaşma yeteneği olarak. Bu nedenle zeki bir insan insanlara yardım etmekte çok iyi olabilir veya insanları incitmede çok iyi olabilir. 7. bölümde hedefler konusunu inceleyeceğiz. Hedeflerle ilgili olarak, aynı zamanda hedeflerinden bahsettiğimizin inceliklerini de açıklamamız gerekiyor. Gelecekteki yepyeni robotik kişisel asistanınızın kendi başına hiçbir hedefi olmadığını, ancak ondan ne yapmasını istiyorsanız onu yapacak ve ondan mükemmel bir İtalyan yemeği pişirmesini istediğinizi varsayalım. İnternete girip İtalyan yemek tariflerini, en yakın markete nasıl gideceğini, makarnanın nasıl süzüldüğünü vb. Araştırırsa, ve sonra malzemeleri başarılı bir şekilde satın alır ve lezzetli bir yemek hazırlarsa, asıl hedef sizin olsa bile muhtemelen bunun akıllı olduğunu düşüneceksiniz. Aslında, talebinizi yaptıktan sonra hedefinizi benimsedi ve ardından kasiyere ödeme yapmaktan Parmesan'ı ızgara yapmaya kadar kendi alt hedefler hiyerarşisine böldü. Bu anlamda, zeki davranış amaca ulaşmakla amansız bir şekilde bağlantılıdır.



Şekil 2.2: Hans Moravec'in, yüksekliğin bilgisayarlar için zorluğu temsil ettiği ve yükselen deniz seviyesinin bilgisayarların yapabildiklerini temsil ettiği "insan yeterliliği manzarası" nın resmi.

Görevlerin zorluğunu, biz insanlar için bunları gerçekleştirmenin ne kadar zor olduğuna göre derecelendirmemiz doğaldır. şekil 2.1 . Ancak bu, bilgisayarlar için ne kadar zor olduklarına dair yanıltıcı bir resim verebilir. 314.159'u ile çarpmak çok daha zor geliyor
271.828, bir fotoğrafı bir arkadaşı tanımaktan çok daha fazla, ancak bilgisayarlar ben doğmadan çok önce bizi aritmetikte parlattı, insan seviyesinde görüntü tanıma ise ancak yakın zamanda mümkün hale geldi. Düşük seviyeli sensorimotor görevlerin, muazzam hesaplama kaynakları gerektirmesine rağmen kolay göründüğü bu gerçeği, Moravec paradoksu olarak bilinir ve beynimizin bu tür görevleri onlara büyük miktarlarda özelleştirilmiş donanım tahsis ederek kolay hissettirmesi gerçeğiyle açıklanır - aslında beynimiz.

Hans Moravec'in bu metaforuna bayılıyorum ve bunu açıklamak için özgürdüm. şekil 2.2 :

> Bilgisayarlar evrensel makinelerdir ve potansiyelleri sınırsız bir görev genişliğine eşit olarak yayılır. İnsan potansiyelleri,

> Öte yandan, hayatta kalmak için uzun süre önemli olan alanlarda güçlüdür, ancak çok uzak şeylerde zayıftır. "Aritmetik" ve "ezberci ezberleme" gibi etiketler içeren alçak arazilere, "teoremi kanıtlayan" ve "satranç oynama" gibi tepelere ve "hareket", "el-göz koordinasyonu" etiketli yüksek dağ zirvelerine sahip bir "insan yeterliliği manzarası" hayal edin. ve "sosyal etkileşim". Bilgisayar performansını geliştirmek, manzarayı yavaşça sular altında bırakan su gibidir. Yarım asır önce, insan hesap makinelerini ve kayıt memurlarını kovarak ovaları boğmaya başladı, ancak çoğumuzu kuru bıraktı. Şimdi sel dağ eteklerine ulaştı ve oradaki karakollarımız geri çekilmeyi düşünüyor. Zirvelerimizde kendimizi güvende hissediyoruz, ancak şu anda onlar da yarım yüzyıl içinde sular altında kalacak. O gün yaklaşırken Arks inşa etmemizi ve denizcilik hayatını benimsememizi öneriyorum! [2]

Bu pasajları yazdığından beri geçen on yıllar boyunca, deniz seviyesi, tahmin ettiği gibi, steroidler üzerindeki küresel ısınma gibi, amansız bir şekilde yükselmeye devam etti ve bazı tepeleri (satranç dahil) çoktan sular altında kaldı. Bundan sonra ne gelecek ve bununla ilgili ne yapmamız gerektiği bu kitabın geri kalanının konusudur.

Deniz seviyesi yükselmeye devam ederken, bir gün devrilme noktasına ulaşarak dramatik bir değişikliği tetikleyebilir. Bu kritik deniz seviyesi, makinelerin yapay zeka tasarımını gerçekleştirebilmelerine karşılık gelen seviyedir. Bu devrilme noktasına ulaşılmadan önce, deniz seviyesinin yükselmesine neden olur *insanlar* makineleri geliştirmek; daha sonra yükselişe neden olabilir *makineler* İnsanların yapabileceğinden potansiyel olarak çok daha hızlı makineleri geliştirmek, hızla tüm karayı sular altında bırakmak. Bu büyüleyici ve tartışmalı fikridir. *tekillik* 4. bölümde keşfederken eğleneceğiz.

Bilgisayar öncüsü Alan Turing, bir bilgisayar belirli bir minimum işlem setini gerçekleştirebilirse, yeterli zaman ve bellek verildiğinde, herhangi bir şeyi yapacak şekilde programlanabileceğini kanıtladı. *hiç* diğer bilgisayar yapabilir. Bu kritik eşiği aşan makinelere *evrensel bilgisayarlar (* aka Turing-evrensel bilgisayarlar); günümüzün tüm akıllı telefonları ve dizüstü bilgisayarları bu anlamda evrenseldir. Benzer bir şekilde, yapay zeka tasarımı için gereken kritik zeka eşiğini, bunun için eşik olarak düşünmeyi seviyorum. *evrensel zeka:*

Yeterli zaman ve kaynak verildiğinde, herhangi bir hedefi gerçekleştirmenin yanı sıra kendisini de başarabilir *hiç* diğer akıllı varlık. Örneğin, daha iyi sosyal beceriler, tahmin becerileri veya yapay zeka tasarım becerileri istediğine karar verirse, bunları elde edebilir. Eğer o

bir robot fabrikasının nasıl kurulacağını bulmaya karar verir, sonra bunu yapabilir. Diğer bir deyişle, evrensel zeka, Life 3.0'a dönüşme potansiyeline sahiptir.

Yapay zeka araştırmacıları arasındaki geleneksel görüş, zekanın sonuçta et, kan veya karbon atomları ile değil, tamamen bilgi ve hesaplamayla ilgili olduğudur. Bu, makinelerin bir gün en azından bizim kadar akıllı olamamasının temel bir nedeni olmadığı anlamına gelir.

Ama fizik bize temel düzeyde her şeyin basitçe madde ve enerjinin hareket ettiğini öğrettiği düşünüldüğünde, gerçekte bilgi ve hesaplama nedir? Bilgi ve hesaplama gibi soyut, soyut ve ruhani bir şey nasıl somut fiziksel şeyler tarafından somutlaştırılabilir? Özellikle fizik yasalarına göre hareket eden bir grup aptal parçacık, zeki dediğimiz davranışları nasıl sergileyebilir?

Bu sorunun cevabının açık olduğunu düşünüyorsanız ve makinelerin bu yüzyılda insanlar kadar zeki olabileceğini düşünürseniz - örneğin bir YZ araştırmacısı olduğunuz için - lütfen bu bölümün geri kalanını atlayın ve doğrudan 3. bölüme geçin. Aksi takdirde, sonraki üç bölümü sizin için özel olarak yazdığımı bilmekten memnuniyet duyacaksınız.

Hafıza Nedir?

Bir atlasın içerdiğini söylersek *bilgi* Dünyayla ilgili olarak, kitabın durumu (özellikle harflere ve görüntülere renklerini veren belirli moleküllerin konumları) ile dünyanın durumu (örneğin, kıtaların konumları) arasında bir ilişki olduğunu kastediyoruz. Kıtalar farklı yerlerde olsaydı o moleküller de farklı yerlerde olurdu. Biz insanlar, kitaplardan ve beyinlere ve sabit disklere kadar bilgi depolamak için bir dizi farklı cihaz kullanıyoruz ve hepsi bu özelliği paylaşıyor: durumlarının, ilgilendiğimiz diğer şeylerin durumuyla ilgili olabileceği (ve bu nedenle bizi bilgilendirebileceği) .

Bunların hepsinin bellek aygıtı olarak kullanışlı olmasını sağlayan ortak hangi temel fiziksel özellikleri var, yani bilgi depolamak için aygıtlar? Cevap hepsinin *birçok farklı uzun ömürlü durumda olabilir*— gerekli olana kadar bilgiyi kodlayacak kadar uzun ömürlüdür. Basit bir örnek olarak, on altı farklı vadiye sahip tepelik bir yüzeye bir top yerleştirdiğinizi varsayalım.

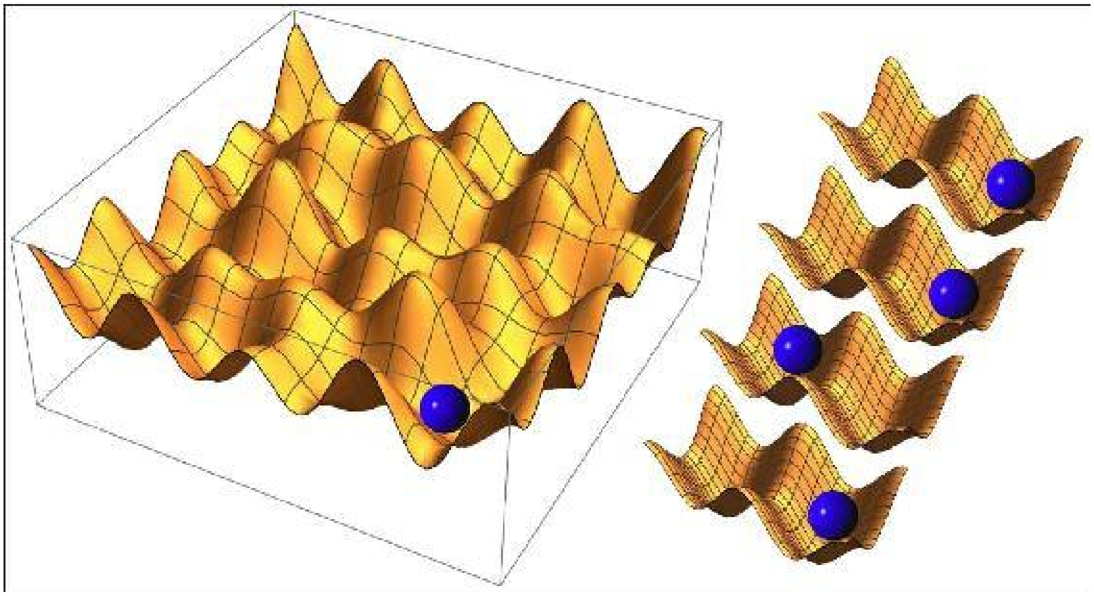
şekil 2.3 . Top yuvarlanıp durduğunda, on altı yerden birinde olacaktır, böylece konumunu 1 ile 16 arasındaki herhangi bir sayıyı hatırlamanın bir yolu olarak kullanabilirsiniz.

Bu hafıza cihazı oldukça sağlamdır, çünkü dış kuvvetler tarafından biraz sallanıp rahatsız edilse bile, top muhtemelen koyduğunuz vadide kalacaktır, böylece hangi numaranın saklandığını hala anlayabilirsiniz. Bu hafızanın bu kadar kararlı olmasının nedeni, topu vadinin dışına çıkarmak, rastgele karışıklıkların sağlayabileceğinden daha fazla enerji gerektirmesidir. Aynı fikir, hareketli bir top için olduğundan çok daha genel olarak sabit anılar sağlayabilir: karmaşık bir fiziksel sistemin enerjisi, her türlü mekanik, kimyasal, elektriksel ve manyetik özelliğe bağlı olabilir ve sistemi değiştirmek için enerji gerektirdiği sürece hatırlamasını istediğiniz durumdan, bu durum kararlı olacaktır. Katıların birçok uzun ömürlü durumu olmasının nedeni budur, halbuki sıvılar ve gazlar böyle değildir: birinin adını altın bir yüzüğün üzerine kazıdığınızda,

Olası en basit bellek aygıtının yalnızca iki kararlı durumu vardır (şekil 2.3).

Bu nedenle, bunu ikili bir rakamı (“bit” kısaltılmış), yani sıfır veya bir kodlaması olarak düşünebiliriz. Daha karmaşık herhangi bir bellek cihazı tarafından saklanan bilgiler, aynı şekilde birden çok bitte depolanabilir: örneğin, birlikte alındığında, gösterilen dört bit şekil 2.3 2 × 2 × 2 × 2 = 16 farklı durumda olabilir 0000, 0001,

0010, 0011,…, 1111, bu nedenle toplu olarak daha karmaşık 16 durumlu sistemle tam olarak aynı bellek kapasitesine sahipler. Bu nedenle, bitleri bilgi atomları olarak düşünebiliriz - daha fazla alt bölümlere ayrılamayan, herhangi bir bilgiyi oluşturmak için birleştirilebilen en küçük bölünmez bilgi yığını. Örneğin, "kelime" kelimesini yazdım ve dizüstü bilgisayarım onu hafızasında 4-sayı dizisi olarak temsil etti ve bu sayıların her birini 8 bit olarak sakladı (her küçük harfi 96 olan bir sayı ile temsil eder) artı alfabedeki sırası). Vurur vurmaz *w* klavyemde tuşa basıldığında, dizüstü bilgisayarım bir *w* ekranımda ve bu görüntü de bitlerle temsil ediliyor: 32 bit, ekranın milyonlarca pikselinin her birinin rengini belirler.



Şekil 2.3: Fiziksel bir nesne, birçok farklı kararlılıkta olabiliyorsa, kullanışlı bir bellek cihazıdır devletler. Soldaki top dört bitlik bilgi etiketlemesini kodlayabilir, hangisi 2 4 = İçinde 16 vadi var. Sağdaki dört top aynı zamanda dört bit bilgiyi de kodluyor - her biri bir bit.

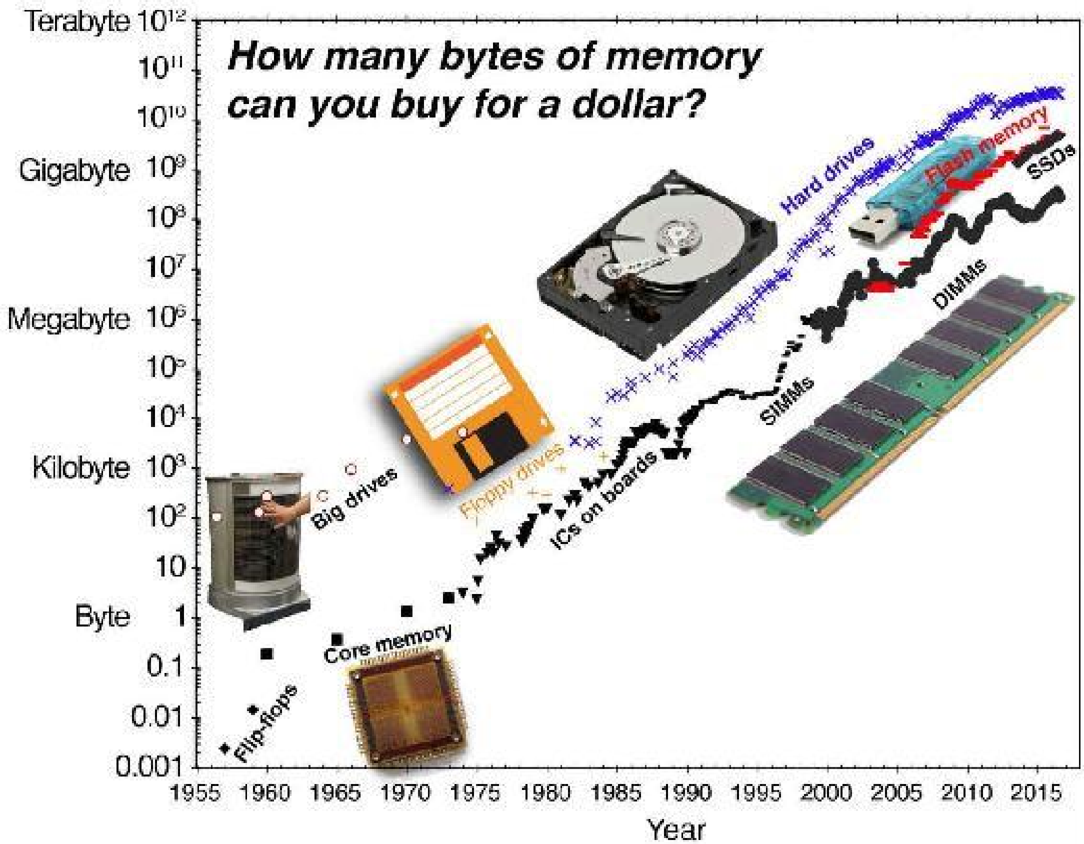
İki durumlu sistemlerin üretimi ve çalışması kolay olduğundan, çoğu modern bilgisayar bilgilerini bit olarak depolar, ancak bu bitler çok çeşitli şekillerde somutlaştırılmıştır. Bir DVD'de, her bit, plastik yüzeyde belirli bir noktada mikroskobik bir çukur olup olmadığına karşılık gelir. Bir sabit sürücüde, her bit, iki yoldan biriyle mıknatıslanan yüzeydeki bir noktaya karşılık gelir. Dizüstü bilgisayarımın çalışma belleğinde, her bit belirli elektronların pozisyonlarına karşılık gelir ve mikro kapasitör adı verilen bir cihazın yüklü olup olmadığını belirler. Işık hızında bile bazı bit türlerinin taşınması uygundur: örneğin, e-postanızı ileten bir optik fiberde, her bit, belirli bir zamanda güçlü veya zayıf bir lazer ışınına karşılık gelir.

Mühendisler, bitleri yalnızca kararlı ve okunması kolay (altın bir yüzük olarak) değil, aynı zamanda yazması da kolay olan sistemlere kodlamayı tercih eder: sabit sürücünüzün durumunu değiştirmek altın kazımaktan çok daha az enerji gerektirir. Ayrıca çalışmaya uygun ve toplu üretime ucuz olan sistemleri tercih ediyorlar. Ancak bunun dışında, bitlerin fiziksel nesneler olarak nasıl temsil edildiğini umursamıyorlar - ve çoğu zaman siz de öyle, çünkü önemli değil! Eğer

arkadaşınıza yazdırması için bir belge e-postayla gönderirseniz, bilgiler sabit sürücünüzdeki mıknatıslamalardan bilgisayarınızın çalışan belleğindeki elektrik yüklerine, kablosuz ağınızdaki radyo dalgalarına, yönlendiricinizdeki voltajlara, bir optik fiberdeki lazer darbelerine kadar hızlı bir şekilde kopyalanabilir. ve son olarak bir kağıt parçası üzerindeki moleküller. Başka bir deyişle, *bilgi, fiziksel alt tabakasından bağımsız olarak kendi başına bir yaşam sürebilir!* Aslında, ilgilendiğimiz şey genellikle bilginin alt tabakadan bağımsız yönüdür: eğer arkadaşınız gönderdiğiniz belgeyi tartışmak için sizi ararsa, muhtemelen voltajlar veya moleküller hakkında konuşmak için aramıyordur. Bu, zeka kadar soyut bir şeyin somut fiziksel şeylerde nasıl somutlaştırılabileceğine dair ilk ipucumuzdur ve yakında bu substrat bağımsızlığı fikrinin sadece bilgi değil, aynı zamanda hesaplama ve öğrenme de dahil olmak üzere nasıl daha derin olduğunu göreceğiz.

Bu substrat bağımsızlığı nedeniyle, akıllı mühendisler, yazılımımızda herhangi bir değişiklik gerektirmeden, bilgisayarlarımızın içindeki bellek cihazlarını, yeni teknolojilere dayalı olarak, önemli ölçüde daha iyi olanlarla defalarca değiştirebildiler. Sonuç, aşağıda gösterildiği gibi muhteşem oldu
şekil 2.4 : Son altmış yılda, bilgisayar belleği her iki yılda bir kabaca yarı yarıya pahalı hale geldi. Sabit diskler 100 milyon kat daha ucuz hale geldi ve sadece depolama yerine hesaplama için yararlı olan daha hızlı anılar, 10 trilyon kat daha ucuz hale geldi. Tüm alışverişlerinizde böyle bir "% 99.99999999999 indirim" alabilseydiniz, New York City'deki tüm gayrimenkulleri yaklaşık 10 sente ve şimdiye kadar yaklaşık bir dolara çıkarılmış olan tüm altını satın alabilirsiniz.

Çoğumuz için hafıza teknolojisindeki olağanüstü gelişmeler kişisel hikayelerle birlikte gelir. 16 kilobayt hafızalı bir bilgisayar için ödeme yapmak için lisede bir şekerci dükkanında çalıştığımı sevgiyle hatırlıyorum ve bunun için lise sınıf arkadaşım Magnus Bodin ile bir kelime işlemci yapıp sattığımda hepsini yazmak zorunda kaldık. işlemesi gereken sözcükler için yeterli bellek bırakacak ultra kompakt makine kodu. 70kB depolayan disketlere alıştıktan sonra, büyük bir 1,44MB depolayabilen ve bütün bir kitabı alabilen daha küçük 3,5 inçlik disketler ve ardından 10MB depolayan ilk sabit diskim tarafından hayrete düştüm. bugünün şarkı indirmelerinden biri. Ergenliğimden kalan bu anılar, geçen gün, 300.000 kat daha fazla kapasiteye sahip bir sabit diske yaklaşık 100 $ harcadığımda neredeyse gerçek dışı hissettirdi.



Şekil 2.4: Son altmış yılda, bilgisayar belleği her iki yılda bir kabaca iki kat daha ucuz hale geldi, bu da kabaca her yirmi yılda bir bin kat daha ucuza denk geliyor. Bir bayt sekiz bite eşittir. John McCallum'un izniyle, http://www.jcmit.net/memoryprice.htm.

İnsanlar tarafından tasarlanmaktan çok gelişen bellek aygıtları ne olacak? Biyologlar, planlarını nesiller arasında kopyalayan ilk yaşam formunun ne olduğunu henüz bilmiyorlar, ancak oldukça küçük olabilir. Cambridge Üniversitesi'nden Philipp Holliger liderliğindeki bir ekip, 2016'da 412 bitlik genetik bilgiyi kodlayan ve kendisinden daha uzun RNA ipliklerini kopyalayabilen bir RNA molekülü yaptı, bu da erken Dünya yaşamının kısa kendi kendini kopyalayan RNA'yı içerdiğine dair "RNA dünyası" hipotezini güçlendirdi. parçacıklar. Şimdiye kadar, doğada evrimleştiği ve kullanıldığı bilinen en küçük bellek cihazı, yaklaşık 40 kilobayt depolayan Candidatus Carsonella ruddii bakterisinin genomudur, oysa insan DNA'mız indirilen bir filme kıyasla yaklaşık 1,6 gigabayt depolar. Gibi

son bölümde bahsettiğimiz gibi, beyinlerimiz genlerimizden çok daha fazla bilgi depolar: 10 gigabaytlık (herhangi bir anda 100 milyar nöronunuzdan hangisinin ateşlediğini belirtir) ve kimyasal / biyolojik olarak 100 terabaytlık (nöronların ne kadar güçlü farklı olduğunu belirtir) sinapslarla bağlanır). Bu sayıları makine hafızalarıyla karşılaştırmak, dünyanın en iyi bilgisayarlarının artık herhangi bir biyolojik sistemi geride bırakabildiğini gösteriyor - bu maliyet hızla düşüyor ve 2016'da birkaç bin dolardı.

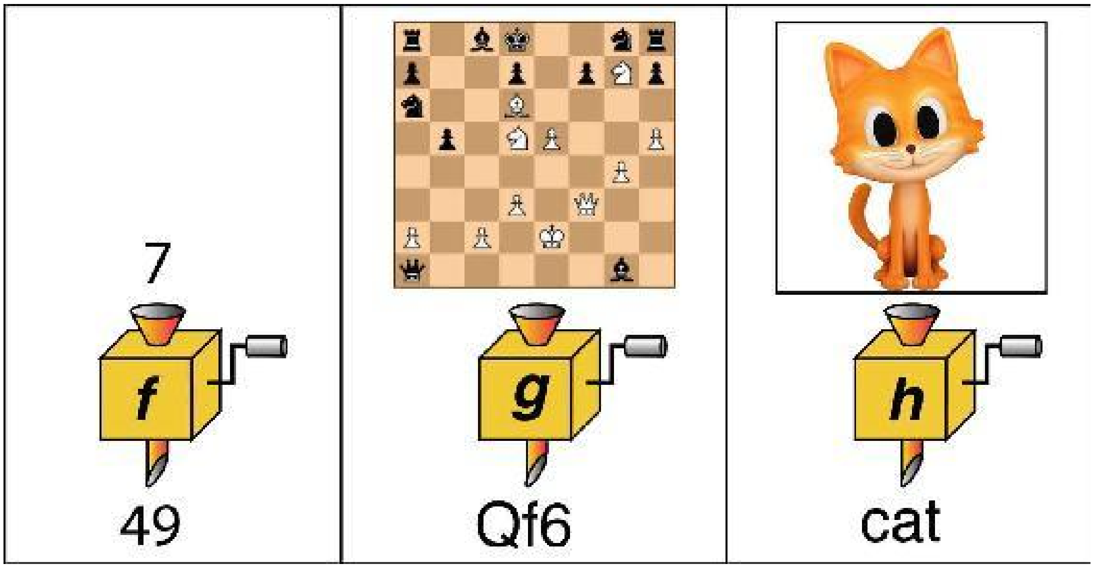
Beyninizdeki bellek, yalnızca nasıl oluşturulduğu açısından değil, aynı zamanda nasıl kullanıldığı açısından da bilgisayar belleğinden çok farklı çalışır. Bir bilgisayardan veya sabit sürücüden anıları belirleyerek alırken *nerede* saklanır, hakkında bir şey belirleyerek beyninizden anıları alırsınız *ne* saklanır. Bilgisayarınızın belleğindeki her bit grubunun sayısal bir adresi vardır ve bir bilgi parçasını almak için bilgisayar, sanki size "Kitaplığıma git, sağdan beşinci kitabı al" derim gibi bakması gereken adresi belirler. üst raf ve bana 314. sayfada ne yazdığını söyle. " Aksine, bir arama motorundan nasıl aldığınıza benzer şekilde beyninizden bilgi alırsınız: bilginin bir parçasını veya bununla ilgili bir şeyi belirtirsiniz ve açılır. Size "ol ya da olmama" dersem ya da Google'da aratırsam, muhtemelen "Olmak ya da olmamak, soru budur." Aslında, alıntının başka bir bölümünü kullansam veya işleri biraz karıştırsam bile muhtemelen işe yarayacaktır. Bu tür bellek sistemleri denir *otomatik çağrışımlı* çünkü adres yerine ilişkilendirmeye göre hatırlıyorlar.

Fizikçi John Hopfield, 1982 tarihli ünlü bir makalesinde, birbirine bağlı nöronlardan oluşan bir ağın kendiliğinden ilişkili bir bellek olarak nasıl işlev görebileceğini gösterdi. Temel fikri çok güzel buluyorum ve çoklu kararlı durumlara sahip herhangi bir fiziksel sistem için işe yarıyor. Örneğin, iki çukuru olan bir yüzeydeki bir top düşünün. şekil 2.3 ve yüzeyi şekillendirelim ki x- Topun durabileceği iki minimanın koordinatları $x = \sqrt{2} \approx 1.41421$ ve $x = \pi$ Sırasıyla ≈ 3,14159. Sadece bunu hatırlarsan π 3'e yakınsa topu $x = 3$ ve daha kesin bir şekilde ortaya çıkmasını izleyin π- değeri en yakın minimuma inerken. Hopfield, karmaşık bir nöron ağının, sistemin yerleşebileceği çok sayıda enerji minimumuna sahip benzer bir manzara sağladığını fark etti ve daha sonra, büyük bir kafa karışıklığına neden olmadan her bin nöron için 138 farklı anıyı sıkıştırabileceğiniz kanıtlandı.

Hesaplama Nedir?

Şimdi fiziksel bir nesnenin bilgiyi nasıl hatırladığını gördük. Ama nasıl hesaplayabilir?

Hesaplama, bir bellek durumunun diğerine dönüştürülmesidir. Başka bir deyişle, bir hesaplama bilgiyi alır ve dönüştürür, matematikçilerin dediği şeyi uygular. *işlevi.* Bilgi için bir kıyma makinesi işlevi görüyorum. şekil 2.5 : Bilgileri en üste koyarsınız, krankı çevirirsiniz ve işlenmiş bilgiyi altta alırsınız - ve bunu farklı girişlerle istediğiniz kadar tekrarlayabilirsiniz. Bu bilgi işleme, aynı girdi ile tekrarlarsanız, her seferinde aynı çıktıyı elde etmeniz anlamında belirleyicidir.

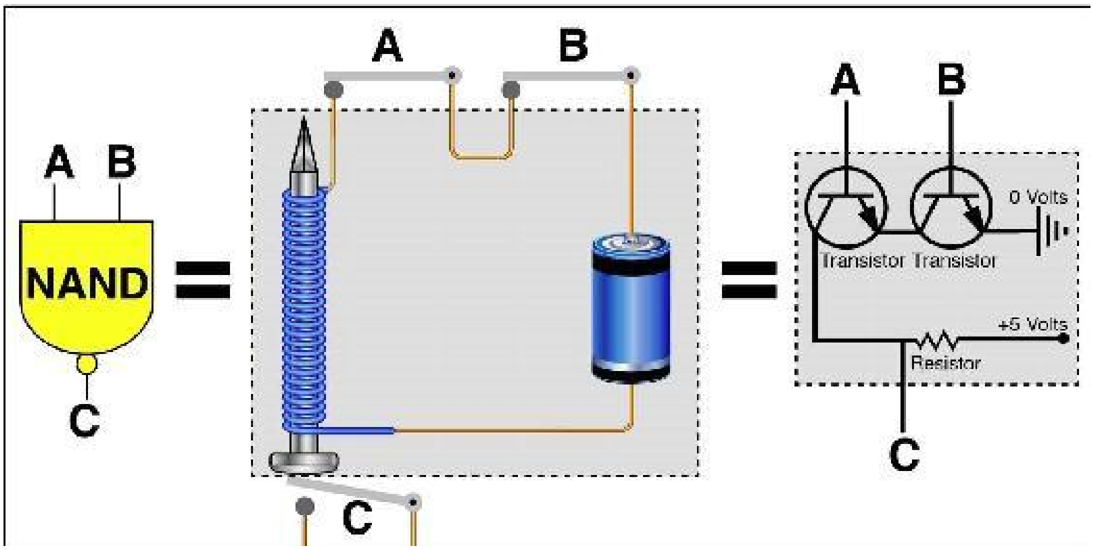


Şekil 2.5: A *hesaplama* bilgiyi alır ve dönüştürür, matematikçilerin dediği şeyi uygular. *işlevi*. İşlev *f (* sol) bir sayıyı temsil eden bitleri alır ve karesini hesaplar. İşlev *g (* orta) bir satranç konumunu temsil eden bitleri alır ve Beyaz için en iyi hamleyi hesaplar. İşlev *h (* sağ) bir görüntüyü temsil eden bitleri alır ve onu açıklayan bir metin etiketini hesaplar.

Kulağa aldatıcı derecede basit gelse de, bu işlev fikri inanılmaz derecede geneldir. Arananlar gibi bazı işlevler oldukça önemsizdir. *DEĞİL* Bu, tek bir biti girer ve tersini verir, böylece sıfırı bire çevirir ve bunun tersi de geçerlidir. Okulda öğrendiğimiz işlevler tipik olarak bir cep hesap makinesindeki düğmelere karşılık gelir, bir veya daha fazla sayı girer ve tek bir sayı çıkarır.

- örneğin, işlev x^2 basitçe bir sayı girer ve kendisiyle çarpılarak çıktı verir. Diğer işlevler son derece karmaşık olabilir. Örneğin, rastgele bir satranç pozisyonunu temsil eden bitleri girecek ve bir sonraki olası en iyi hamleyi temsil eden bitleri çıkaracak bir işleve sahipseniz, bunu Dünya Bilgisayar Satranç Şampiyonası'nı kazanmak için kullanabilirsiniz. Dünyanın tüm finansal verilerini giren ve satın alınabilecek en iyi hisse senetlerini çıkaran bir işleve sahipseniz, yakında son derece zengin olacaksınız. Birçok AI araştırmacısı, kariyerlerini belirli işlevlerin nasıl uygulanacağını bulmaya adamıştır. Örneğin, makine çevirisi araştırmasının amacı, bir dildeki metni temsil eden bitleri girerek ve aynı metni başka bir dilde temsil eden bitleri çıkaran bir işlev uygulamaktır ve otomatik altyazı koyma araştırmasının amacı, giriş yapmaktır.

bir görüntüyü temsil eden ve onu açıklayan metni temsil eden çıktı bitleri (şekil 2.5).



Şekil 2.6: Sözde NAND geçidi, A = B = 1 ve C = 1 ise C = 0 kuralına göre, giriş olarak A ve B olmak üzere iki bit alır ve çıkış olarak bir bit C'yi hesaplar. Birçok fiziksel sistem NAND kapıları olarak kullanılabilir. Ortadaki örnekte, anahtarlar 0 = açık, 1 = kapalı olan bitler olarak yorumlanır ve A ve B anahtarlarının her ikisi de kapalı olduğunda, bir elektromıknatıs anahtarı açar
C. En sağdaki örnekte, voltajlar (elektriksel potansiyeller) 1 = beş volt, 0 = sıfır volt olarak yorumlanır ve A ve B kablolarının her ikisi de beş voltta olduğunda, iki transistör elektrik iletir ve C teli düşer. yaklaşık sıfır volt.

Başka bir deyişle, oldukça karmaşık işlevleri uygulayabilirseniz, son derece karmaşık hedefleri gerçekleştirebilen akıllı bir makine oluşturabilirsiniz. Bu, maddenin nasıl zeki olabileceğine dair sorumuzu daha keskin bir odak noktasına getiriyor: özellikle, görünüşte aptal görünen bir madde yığını karmaşık bir işlevi nasıl hesaplayabilir?

Altın bir yüzük veya başka bir statik bellek cihazı olarak hareketsiz kalmak yerine, karmaşıklık sergilemelidir. *dinamikler* böylece gelecekteki durumu mevcut duruma bazı karmaşık (ve umarız kontrol edilebilir / programlanabilir) bir şekilde bağlıdır. Atom düzeni, ilginç hiçbir şeyin değişmediği sert bir katıdan daha az düzenli olmalı, ancak bir sıvı veya gazdan daha düzenli olmalıdır. Spesifik olarak, sistemin, girdi bilgilerini kodlayan bir duruma koyarsak, bir süre fizik yasalarına göre evrimleşmesine izin vermesi ve ardından ortaya çıkan son durumu çıktı bilgisi olarak yorumlama özelliğine sahip olmasını istiyoruz. , o zaman çıktı, girişin istenen işlevidir. Eğer durum buysa, diyebiliriz

sistemimizin işlevimizi hesapladığı.

Bu fikrin ilk örneği olarak, çok basit bir şeyi nasıl oluşturabileceğimizi inceleyelim.
(ama aynı zamanda çok önemli) işlev olarak adlandırılan *NAND kapısı* *3 eski aptal meselenin dışında. Bu fonksiyon iki bit girdi ve bir bit çıktı: eğer her iki giriş de 1 ise 0 çıktı; diğer tüm durumlarda, çıkış verir 1. İki anahtarı seri olarak bir batarya ve bir elektromıknatısla bağlarsak, elektromıknatıs yalnızca ilk anahtar *ve* ikinci anahtar kapalıdır ("açık"). Elektromıknatısın altına üçüncü bir anahtar yerleştirelim. şekil 2.6 , öyle ki mıknatıs her açıldığında onu çekerek açacaktır. İlk iki anahtarı giriş bitleri ve üçüncüsünü çıkış biti olarak yorumlarsak (0 = anahtar açık ve 1 = anahtar kapalı), o zaman kendimize bir NAND geçidimiz olur: üçüncü anahtar yalnızca ilk anahtar ise açıktır ikisi kapalı. NAND kapıları oluşturmanın daha pratik olan birçok başka yolu vardır - örneğin, aşağıdaki gibi transistörleri kullanmak

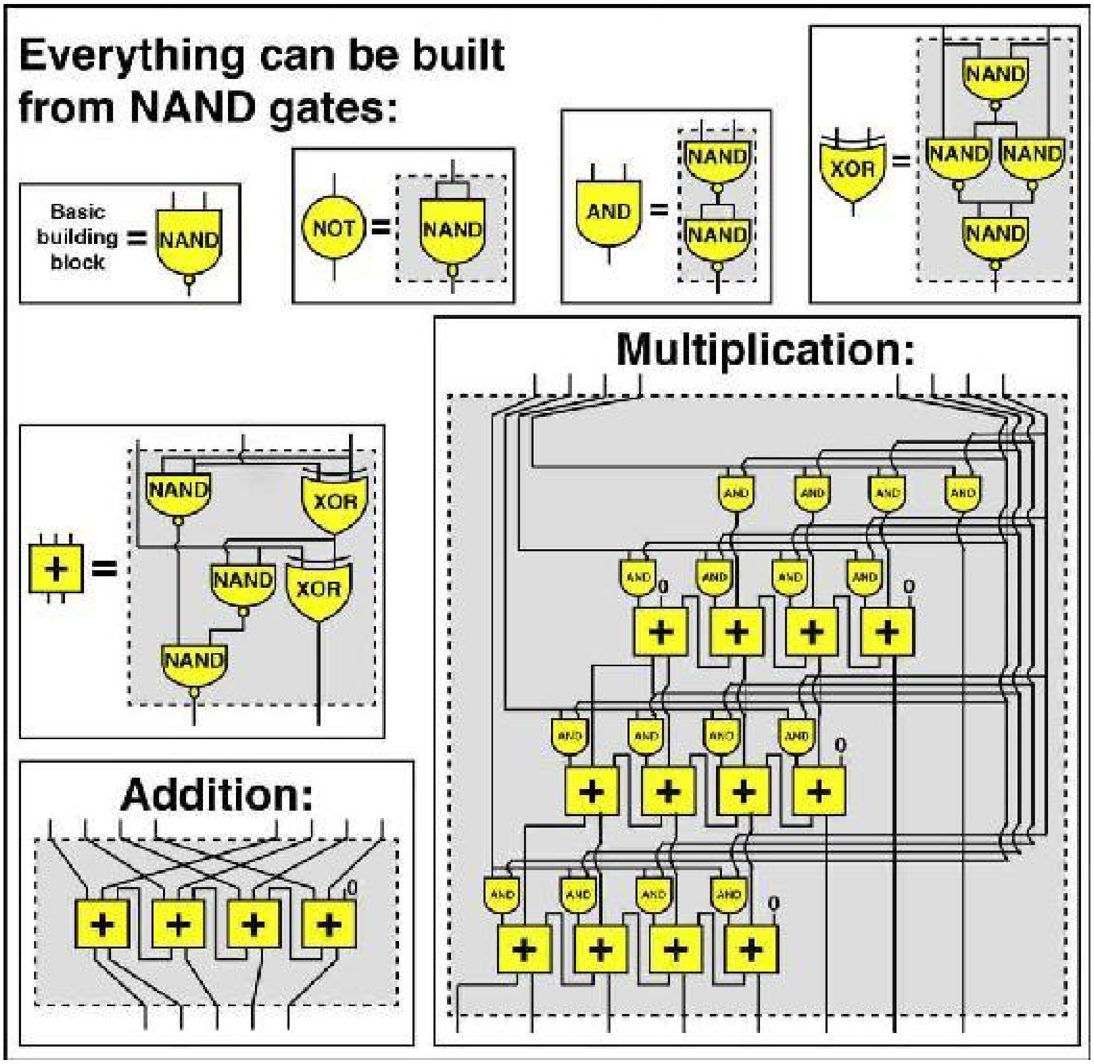
şekil 2.6 . Günümüz bilgisayarlarında, NAND kapıları tipik olarak mikroskobik transistörlerden ve silikon levhalara otomatik olarak kazınabilen diğer bileşenlerden yapılmıştır.

Bilgisayar biliminde, NAND geçitlerinin *evrensel,* uygulayabileceğiniz anlamına gelir *hiç* iyi tanımlanmış işlev basitçe
NAND kapılarını birbirine bağlayarak. * 4 Dolayısıyla, yeterince NAND geçidi inşa edebilirseniz, her şeyi hesaplayan bir cihaz oluşturabilirsiniz! Bunun nasıl çalıştığını tatmak isterseniz, şekil 2.7 NAND geçitlerinden başka hiçbir şey kullanmadan sayıları nasıl çarpabilirim.

MIT araştırmacıları Norman Margolus ve Tommaso Toffoli adı icat etti
bilgisayar rastgele hesaplamalar yapabilen herhangi bir madde için. Compronium yapmanın özellikle zor olmak zorunda olmadığını gördük: maddenin sadece istenen herhangi bir şekilde birbirine bağlanan NAND geçitlerini uygulayabilmesi gerekiyor. Gerçekte, sayısız başka bilgisayar türü de vardır. Aynı zamanda işe yarayan basit bir varyant, NAND geçitlerinin, yalnızca her iki giriş de 0 olduğunda 1 çıkan NOR geçitleriyle değiştirilmesini içerir. Bir sonraki bölümde, rasgele hesaplamaları da uygulayabilen, yani computronium gibi davranabilen sinir ağlarını inceleyeceğiz. Bilim adamı ve girişimci Stephen Wolfram, aynı şeyin, komşu bitlerin yaptıklarına göre bitleri tekrar tekrar güncelleyen hücresel otomata adı verilen basit cihazlar için de geçerli olduğunu gösterdi. Zaten 1936'da geri döndüm. bilgisayar öncüsü Alan Turing, bir dönüm noktası niteliğindeki bir makalede, bir bant şeridindeki sembolleri manipüle edebilen basit bir makinenin (artık "evrensel Turing makinesi" olarak bilinir) keyfi hesaplamalar da uygulayabileceğini kanıtladı. Özetle, sadece

Maddenin iyi tanımlanmış herhangi bir hesaplamayı gerçekleştirmesi mümkün mü, ancak bu, birçok farklı yolla mümkündür.

Daha önce de belirtildiği gibi, Turing, 1936 tarihli makalesinde daha da derin bir şeyi kanıtladı: eğer bir bilgisayar türü belirli bir minimum işlem setini gerçekleştirebiliyorsa, o zaman *evrensel* yeterli kaynak verildiği anlamda, başka herhangi bir bilgisayarın yapabileceği her şeyi yapabilir. Turing makinesinin evrensel olduğunu gösterdi ve fiziğe daha yakından bağlanarak, bu evrensel bilgisayarlar ailesinin aynı zamanda bir NAND kapıları ağı ve birbirine bağlı bir nöron ağı kadar çeşitli nesneleri de içerdiğini gördük. Aslında, Stephen Wolfram şunu ileri sürmüştür: *çoğu* Hava sistemlerinden beyinlere kadar önemsiz olmayan fiziksel sistemler, keyfi olarak büyük ve uzun ömürlü yapılabilselerdi evrensel bilgisayarlar olurdu.



Şekil 2.7: *Hiç* İyi tanımlanmış hesaplama, NAND geçitlerinden başka hiçbir şeyi akıllıca birleştirmeyerek gerçekleştirilebilir. Örneğin, her ikisinin üstündeki toplama ve çarpma modülleri, 4 bit ile temsil edilen iki ikili sayı girer ve sırasıyla 5 bit ve 8 bit ile temsil edilen bir ikili sayı çıkarır. Daha küçük modüller NOT, AND, XOR ve + (üç ayrı biti 2 bitlik ikili sayıya toplar) sırayla NAND kapılarından oluşturulur. Bu rakamı tam olarak anlamak son derece zordur ve bu kitabın geri kalanını takip etmek için tamamen gereksizdir; Bunu sadece evrensellik fikrini örneklemek ve içimdeki geekimi tatmin etmek için buraya dahil ediyorum.

Tam olarak aynı hesaplamanın yapılabileceği gerçeği *hiç* evrensel bilgisayar şu anlama gelir *hesaplama substrattan bağımsızdır* bilginin olduğu gibi: bilgisinden bağımsız olarak kendi başına bir hayat sürebilir.

fiziksel substrat! Dolayısıyla, gelecekteki bir bilgisayar oyununda bilinçli bir süper zeki karakter iseniz, bir Windows masaüstünde mi, bir Mac OS dizüstü bilgisayarda mı yoksa bir Android telefonda mı koştuğunuzu bilemezsiniz, çünkü alt tabakadan bağımsız olursunuz. Ayrıca, mikroişlemcinin ne tür transistör kullandığını bilmenin hiçbir yolu olamazdı.

İlk önce bu önemli substrat bağımsızlığı fikrini takdir etmeye başladım çünkü fizikte bunun birçok güzel örneği var. Örneğin dalgaların hız, dalga boyu ve frekans gibi özellikleri vardır ve biz fizikçiler itaat ettikleri denklemleri hangi maddenin içinde dalgalandıklarını bilmeye bile gerek kalmadan inceleyebiliriz. Bir şey duyduğunuzda, ses dalgalarını tespit ediyorsunuz. Hava dediğimiz gazların karışımında zıplayan moleküllerin neden olduğu ve bu dalgalar hakkında her türlü ilginç şeyi hesaplayabiliriz - mesafenin karesi olarak yoğunluklarının nasıl azaldığı, açık kapılardan geçerken nasıl eğildikleri gibi ve nasıl duvarlardan sekip yankılara neden oluyorlar

- havanın neyden yapıldığını bilmeden. Aslında, moleküllerden oluştuğunu bilmemize bile gerek yok: oksijen, nitrojen, karbondioksit vb. Hakkındaki tüm ayrıntıları görmezden gelebiliriz, çünkü dalganın substratının önemli olan ve ünlü dalga denklemine giren tek özelliği ölçebileceğimiz tek bir sayıdır: bu durumda saniyede yaklaşık 300 metre olan dalga hızı. Nitekim geçen bahar MIT öğrencilerime anlattığım bu dalga denklemi, fizikçilerin atomların ve moleküllerin var olduğunu bile belirlemesinden çok önce keşfedildi ve büyük ölçüde kullanıldı!

Bu dalga örneği, üç önemli noktayı göstermektedir. İlk olarak, substrat bağımsızlığı, bir substratın gereksiz olduğu anlamına gelmez, ancak ayrıntılarının çoğunun önemli olmadığı anlamına gelir. Açıkçası, gaz yoksa bir gazda ses dalgalarına sahip olamazsınız, ancak herhangi bir gaz yeterli olacaktır. Benzer şekilde, madde olmadan hesaplama yapamayacağınız açıktır, ancak herhangi bir madde NAND kapılarına, bağlı nöronlara veya evrensel hesaplamayı sağlayan başka bir yapı bloğuna yerleştirilebildiği sürece işe yarar. İkincisi, substrattan bağımsız fenomen, substratından bağımsız olarak kendi başına bir yaşam sürüyor. Bir dalga, su moleküllerinden hiçbiri yapmasa bile, bir spor stadyumunda "dalgayı" yapan taraftarlar gibi çoğunlukla yukarı ve aşağı sallanırlar. Üçüncüsü, genellikle ilgilendiğimiz şey yalnızca substrattan bağımsız yöndür: bir sörfçü genellikle dalganın ayrıntılı moleküler bileşiminden çok konumu ve yüksekliği ile ilgilenir. Bunun bilgi için nasıl doğru olduğunu gördük ve hesaplama için de doğru: eğer iki programcı kodlarında ortaklaşa bir hata arıyorlarsa, muhtemelen transistörleri tartışmıyorlardır.

Şimdi, somut fiziksel şeylerin zeka kadar soyut, soyut ve ruhani bir şeye nasıl yol açabileceğiyle ilgili açılış sorumuzun cevabına vardık: çok fiziksel olmayan bir his veriyor çünkü substrattan bağımsızdır ve bir ömür sürüyor. fiziksel ayrıntılara bağlı olmayan veya bunları yansıtmayan kendine ait. Kısacası, hesaplama, parçacıkların uzay-zaman düzenlemesinde bir kalıptır ve asıl önemli olan parçacıklar değil, kalıptır! Madde önemli değil.

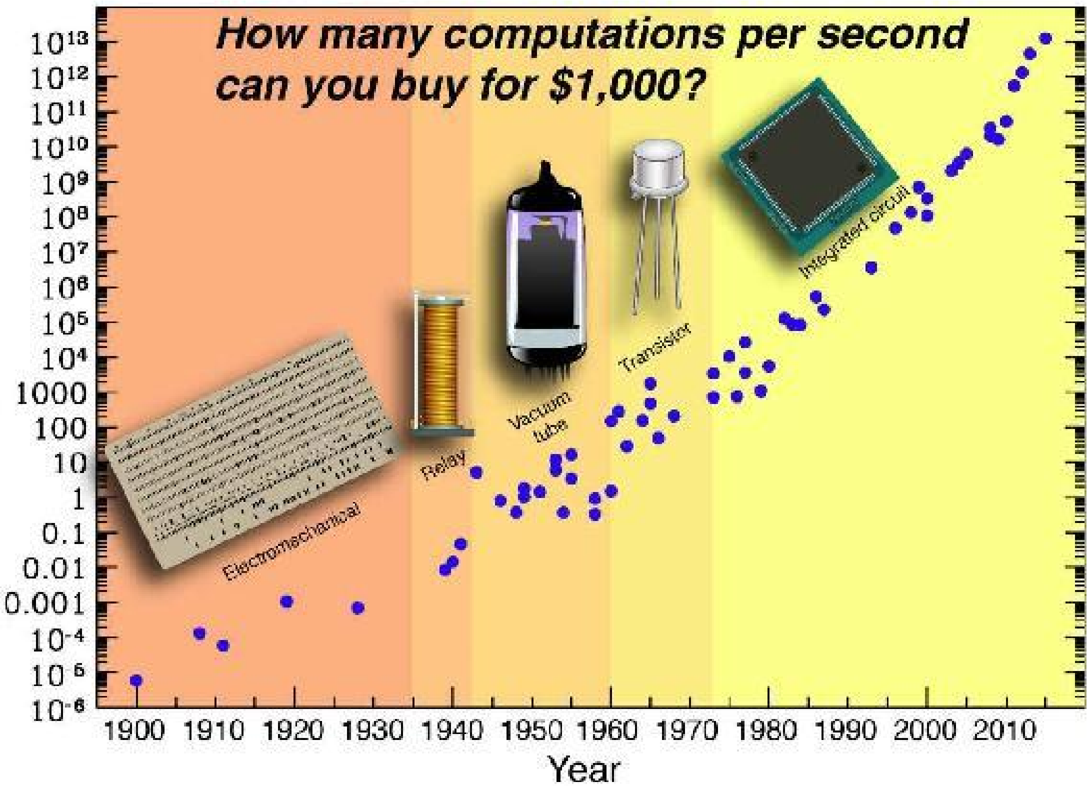
Başka bir deyişle, donanım meseledir ve yazılım da kalıptır. Hesaplamanın bu substrat bağımsızlığı, yapay zekanın mümkün olduğu anlamına gelir: zeka et, kan veya karbon atomu gerektirmez.

Bu substrat bağımsızlığı nedeniyle, zeki mühendisler, yazılımı değiştirmeden bilgisayarlarımızın içindeki teknolojileri defalarca önemli ölçüde daha iyi teknolojilerle değiştirmeyi başardılar. Sonuçlar, bellek aygıtları için olduğu kadar muhteşem oldu. Gösterildiği gibi şekil 2.8 , hesaplama her iki yılda bir kabaca yarı yarıya pahalı hale gelmeye devam ediyor ve bu eğilim şimdi bir yüzyılı aşkın süredir devam etti ve bilgisayar maliyetini düşürdü.

milyon milyon milyon (10^{18}) büyükannelerim doğduğundan beri. Her şey bir milyon milyon kat daha ucuz olsaydı, o zaman yüzde biri bu yıl Dünya'da üretilen tüm mal ve hizmetleri satın almanızı sağlardı. Maliyetlerdeki bu dramatik düşüş, elbette, hesaplamanın bugünlerde her yerde olmasının, geçmiş yılların bina boyutundaki bilgi işlem tesislerinden evlerimize, arabalarımıza ve ceplerimize yayılmasının ve hatta spor ayakkabı gibi beklenmedik yerlerde ortaya çıkmasının temel nedenlerinden biridir.

Neden teknolojimiz düzenli aralıklarla gücünü ikiye katlayarak matematikçilerin üstel büyüme dediği şeyi sergiliyor? Aslında, neden sadece transistör minyatürleştirmesi açısından olmuyor (

Moore yasası), aynı zamanda daha geniş anlamda bir bütün olarak hesaplama için (şekil 2.8), bellek için (şekil 2.4) ve genom dizilemeden beyin görüntülemeye kadar pek çok başka teknoloji için? Ray Kurzweil, bu ısrarlı ikiye katlanma fenomenine "hızlanan geri dönüş yasası" adını veriyor.



Şekil 2.8: 1900'den beri, hesaplama her iki yılda bir kabaca iki kat daha ucuza geldi. Grafik, saniyedeki kayan nokta işlemlerinde (FLOPS) ölçülen bilgi işlem gücünü gösterir.

1.000 $ 'a satın alınabilir. [3] Bir kayan noktayı tanımlayan özel hesaplama

işlem yaklaşık 10'a karşılık gelir 5 bit çevirmeleri veya NAND değerlendirmeleri gibi temel mantıksal işlemler.

Doğada bildiğim tüm ısrarlı ikiye katlama örnekleri aynı temel nedene sahiptir ve bu teknolojik olan bir istisna değildir: her adım bir sonrakini yaratır. Örneğin, gebe kaldıktan hemen sonra üstel büyüme yaşadınız: her bir hücreniz bölündü ve kabaca günde iki hücre oluşturdu, bu da toplam hücre sayınızın her geçen gün 1, 2, 4, 8, 16 gibi artmasına neden oldu. üzerinde. Kozmik kökenlerimizin en popüler bilimsel teorisine göre, *şişirme,* bizim bebek Evrenimiz bir zamanlar sizin yaptığınız gibi katlanarak büyüdü, bir atomdan çok daha küçük ve daha hafif olan bir benek teleskoplarımızla şimdiye kadar gördüğümüz tüm galaksilerden daha büyük büyüyene kadar boyutunu düzenli aralıklarla tekrar tekrar ikiye katladı. Yine, neden, her ikiye katlama adımının bir sonrakine neden olduğu bir süreçti. Teknoloji de böyle ilerliyor: bir kez

teknoloji iki kat daha güçlü hale gelir, genellikle iki kat daha güçlü bir teknoloji tasarlamak ve inşa etmek için kullanılabilir ve Moore yasasının ruhunda tekrarlanan kapasitenin ikiye katlanmasını tetikler.

Teknolojik gücümüzün ikiye katlanması kadar düzenli olarak gerçekleşen bir şey, ikiye katlamanın sona erdiği iddialarının ortaya çıkmasıdır. Evet, Moore yasası elbette sona erecek, yani küçük transistörlerin nasıl yapılacağına dair fiziksel bir sınır var. Ancak bazı insanlar yanlışlıkla Moore yasasının teknolojik gücümüzün ısrarla ikiye katlanmasıyla eş anlamlı olduğunu varsayıyor. Aksine, Ray Kurzweil, Moore yasasının, bilgi işlemde üstel büyüme getiren ilk değil beşinci teknolojik paradigmayı içerdiğine dikkat çeker. şekil 2.8 : Bir teknoloji gelişmeyi bıraktığında, onu daha da iyi biriyle değiştirdik. Artık vakum tüplerimizi küçültmeye devam edemediğimizde, onları transistörlerle ve ardından elektronların iki boyutta hareket ettiği entegre devrelerle değiştirdik. Bu teknoloji sınırlarına ulaştığında, deneyebileceğimiz birçok başka alternatif var - örneğin, üç boyutlu devreler kullanmak ve teklifimizi yapmak için elektronlardan başka bir şey kullanmak.

Bir sonraki gişe rekorları kıran hesaplama altyapısının ne olacağını kimse kesin olarak bilmiyor, ancak fizik yasalarının koyduğu sınırların yakınında olmadığımızı biliyoruz. MIT meslektaşım Seth Lloyd bu temel sınırın ne olduğunu çözdü ve 6. bölümde daha ayrıntılı olarak keşfedeceğimiz gibi, bu sınır bir

33 büyüklük mertebesi (10^{33} Bir madde kümesinin ne kadar hesaplama yapabildiğine dair günümüzün sanatının ötesinde. Dolayısıyla, bilgisayarlarımızın gücünü her iki yılda bir ikiye katlamaya devam etsek bile, bu son sınıra ulaşmamız iki yüzyıldan fazla sürecektir.

Tüm evrensel bilgisayarlar aynı hesaplamaları yapabilse de, bazıları diğerlerinden daha verimlidir. Örneğin, milyonlarca çarpma gerektiren bir hesaplama, olduğu gibi ayrı transistörlerden oluşturulmuş milyonlarca ayrı çarpma modülü gerektirmez. şekil 2.6 : uygun girişlerle arka arkaya birçok kez kullanabildiğinden, bu tür tek bir modüle ihtiyaç duyar. Bu verimlilik ruhunda, çoğu modern bilgisayar, hesaplamaların birden çok zaman adımına bölündüğü ve bu sırada bilgilerin bellek modülleri ve hesaplama modülleri arasında ileri geri karıştırıldığı bir paradigma kullanır. Bu hesaplamalı mimari, Alan Turing, Konrad Zuse, Presper Eckert, John Mauchly ve John von Neumann gibi bilgisayar öncüleri tarafından 1935 ile 1945 yılları arasında geliştirildi. Daha spesifik olarak, bilgisayar belleği hem verileri hem de

yazılım (bir program, yani verilerle ne yapılacağına dair talimatların bir listesi). Her zaman adımında, bir merkezi işlem birimi (CPU), programdaki bir sonraki talimatı yürütür ve bu, verilerin bir kısmına uygulanacak bazı basit işlevleri belirtir. Bilgisayarın bir sonraki adımda ne yapılacağını takip eden kısmı, hafızasının yalnızca başka bir kısmıdır. *program sayıcı,* programda mevcut satır numarasını kaydeder. Bir sonraki talimata gitmek için, program sayacına bir tane eklemeniz yeterlidir. Programın başka bir satırına atlamak için, bu satır numarasını program sayacına kopyalamanız yeterlidir - bu, sözde "if" ifadeleri ve döngülerin uygulanma şeklidir.

Bugünün bilgisayarları genellikle ek hız kazanır *paralel işleme,* Bu, modüllerin bu yeniden kullanımının bir kısmını akıllıca geri alır: eğer bir hesaplama paralel olarak yapılabilecek parçalara bölünebiliyorsa (çünkü bir parçanın girdisi diğerinin çıktısını gerektirmez), o zaman parçalar aynı anda farklı şekilde hesaplanabilir. donanımın parçaları.

Nihai paralel bilgisayar bir *kuantum bilgisayar.* Kuantum hesaplamanın öncüsü David Deutsch, tartışmalı bir şekilde, "kuantum bilgisayarların, çoklu evrende kendilerinin çok sayıda versiyonuyla bilgileri paylaştığını" ve bir anlamda Evrenimizde yanıtları daha hızlı alabileceğini savunuyor.
bu diğer sürümlerden yardım almak. [4] Ticari olarak rekabetçi bir kuantum bilgisayarın önümüzdeki on yıllarda inşa edilip edilemeyeceğini henüz bilmiyoruz, çünkü hem kuantum fiziğinin düşündüğümüz gibi çalışıp çalışmadığına hem de göz korkutucu teknik zorlukların üstesinden gelme yeteneğimize bağlı, ancak çevredeki şirketler ve hükümetler dünya bu olasılığa yılda on milyonlarca dolar bahis yapıyor. Kuantum bilgisayarlar sıradan hesaplamaları hızlandıramasa da, kriptosistemleri kırmak ve sinir ağlarını eğitmek gibi belirli hesaplama türlerini önemli ölçüde hızlandırabilen akıllı algoritmalar geliştirildi. Bir kuantum bilgisayar ayrıca atomlar, moleküller ve yeni malzemeler dahil olmak üzere kuantum mekanik sistemlerin davranışını verimli bir şekilde simüle edebilir.

Öğrenme Nedir?

Bir cep hesap makinesi aritmetik bir yarışmada beni ezebilir, ancak ne kadar pratik yaparsa yapsın hızını veya doğruluğunu asla iyileştirmez. Öğrenmiyor: örneğin, karekök düğmesine her bastığımda, tamamen aynı işlevi tamamen aynı şekilde hesaplıyor. Benzer şekilde, satrançta beni yenen ilk bilgisayar programı hatalarından asla ders çıkarmadı, sadece zeki programcısının iyi bir sonraki hamleyi hesaplamak için tasarladığı bir işlevi uyguladı. Tersine, Magnus Carlsen beş yaşında ilk satranç oyununu kaybettiğinde, onu on sekiz yıl sonra Dünya Satranç Şampiyonu yapan bir öğrenme sürecine başladı.

Öğrenme yeteneği, tartışmasız genel zekanın en büyüleyici yönüdür. Aptal gibi görünen bir madde kümesinin nasıl hatırlayıp hesaplayabildiğini zaten görmüştük, ama nasıl öğrenebilir? Zor bir sorunun cevabını bulmanın bir işlevi hesaplamaya karşılık geldiğini ve uygun şekilde düzenlenmiş maddenin herhangi bir hesaplanabilir işlevi hesaplayabildiğini gördük. Biz insanlar ilk olarak cep hesap makinelerini ve satranç programlarını yarattığımızda, *Biz* düzenleme yaptı. Öğenin öğrenilmesi için, bunun yerine yeniden düzenlenmelidir *kendisi* sadece fizik yasalarına uyarak istenen işlevi hesaplamada daha iyi ve daha iyi hale gelmek.

Öğrenme sürecini aydınlatmak için, önce çok basit bir fiziksel sistemin aşağıdaki basamakları nasıl öğrenebileceğini düşünelim. π ve diğer numaralar. Yukarıda birçok vadinin olduğu bir yüzeyin nasıl olduğunu gördük (bkz. şekil 2.3) bir bellek cihazı olarak kullanılabilir: örneğin, vadilerden birinin tabanı konumundaysa $x = \pi \approx 3{,}14159$ ve yakınlarda başka vadi yoksa, o zaman bir top koyabilirsiniz $x = 3$ ve sistemin, topun aşağıya doğru yuvarlanmasına izin vererek eksik ondalık sayıları hesaplamasını izleyin. Şimdi, yüzeyin yumuşak kilden yapıldığını ve boş bir levha olarak tamamen düz bir şekilde başladığını varsayalım. Bazı matematik meraklıları, topu en sevdikleri sayıların her birinin yerine defalarca yerleştirirse, yerçekimi bu yerlerde kademeli olarak vadiler oluşturacak ve ardından kil yüzeyi bu saklanan anıları hatırlamak için kullanılabilecektir. Başka bir deyişle, kil yüzeyinde *öğrendi* gibi sayıların basamaklarını hesaplamak için π.

Beyin gibi diğer fiziksel sistemler, aynı fikre dayanarak çok daha verimli bir şekilde öğrenebilir. John Hopfield, yukarıda bahsedilen ağının

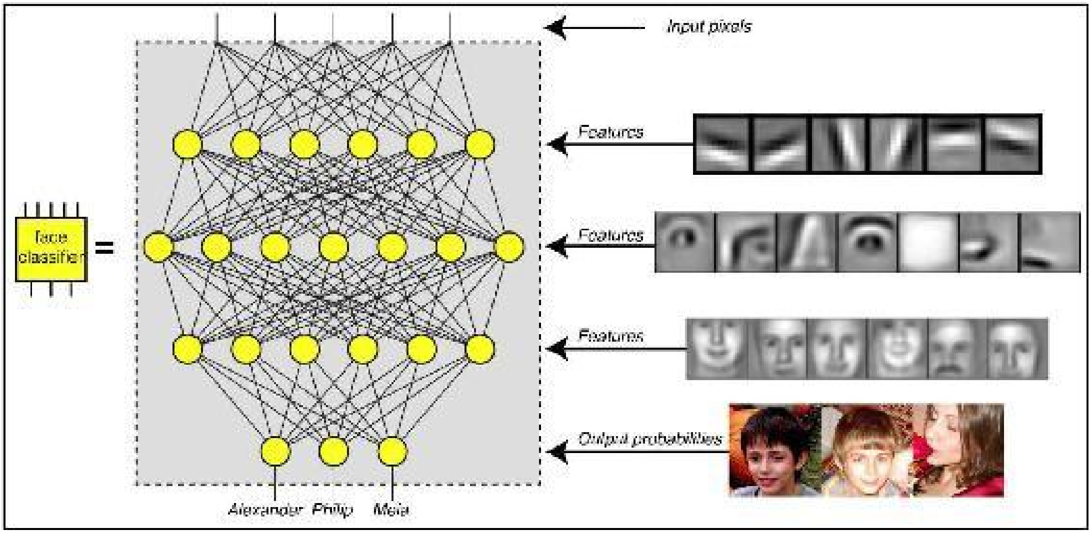
Birbirine bağlı nöronlar benzer bir şekilde öğrenebilirler: eğer onu tekrar tekrar belirli durumlara koyarsanız, bu durumları yavaş yavaş öğrenecek ve yakındaki herhangi bir durumdan onlara geri dönecektir. Aile üyelerinizin her birini birçok kez gördüyseniz, neye benzediklerine dair anılar, onlarla ilgili herhangi bir şey tarafından tetiklenebilir.

Sinir ağları artık hem biyolojik hem de yapay zekayı dönüştürdü ve son zamanlarda AI alt alanına hükmetmeye başladı.
makine öğrenme (deneyim yoluyla gelişen algoritmaların incelenmesi). Bu tür ağların nasıl öğrenebileceğini daha derinlemesine incelemeden önce, önce nasıl hesaplayabileceklerini anlayalım. Bir sinir ağı, basitçe birbirlerinin davranışını etkileyebilen, birbirine bağlı bir grup nörondur. Beyniniz, gökadamızdaki yıldızlar kadar çok sayıda nöron içerir: yüz milyarlık beyzbol sahasında. Ortalama olarak, bu nöronların her biri, adı verilen kavşaklar aracılığıyla yaklaşık bin diğerine bağlanır. *sinapslar,* ve beyninizdeki bilgilerin çoğunu kodlayan bu kabaca yüz trilyon sinaps bağlantısının güçlü yanlarıdır.

Sinapsları temsil eden çizgilerle birbirine bağlanan nöronları temsil eden noktaların bir koleksiyonu olarak şematik olarak bir sinir ağı çizebiliriz (bkz. şekil 2.9). Gerçek dünyadaki nöronlar, bu şematik gösterime hiç benzemeyen çok karmaşık elektrokimyasal cihazlardır: aksonlar ve dendritler gibi isimlerle farklı parçaları içerirler, çok çeşitli şekillerde çalışan birçok farklı nöron türü vardır ve nasıl yapıldığına dair kesin ayrıntılar ve bir nörondaki elektriksel aktivite diğer nöronları etkilediğinde hala aktif çalışmanın konusudur. Bununla birlikte, yapay zeka araştırmacıları, tüm bu karmaşıklıkları göz ardı etse ve gerçek biyolojik nöronları, hepsi aynı olan ve çok basit kurallara uyan son derece basit simüle edilmiş nöronlarla değiştirse bile, sinir ağlarının dikkate değer şekilde birçok karmaşık görevde insan seviyesinde performansa ulaşabileceğini göstermiştir. Şu anda böyle bir model için en popüler model *yapay sinir ağı*

her bir nöronun durumunu tek bir sayı ile ve her sinapsın gücünü tek bir sayı ile temsil eder. Bu modelde, her nöron, normal zaman adımlarında durumunu, bağlı tüm nöronlardan gelen girdilerin ortalamasını alarak, onları sinaptik güçlerle ağırlıklandırarak, isteğe bağlı olarak bir sabit ekleyerek ve ardından *aktivasyon fonksiyonu* hesaplanacak sonuca

bir sonraki durumu. *5 Bir sinir ağını işlev olarak kullanmanın en kolay yolu, onu *ileri besleme,* bilginin yalnızca bir yönde aktığı şekil 2.9 , işleve girdiyi üstte bir nöron katmanına takmak ve çıktıyı altta bir nöron katmanından çıkarmak.



Şekil 2.9: Bir nöron ağı, bir NAND geçit ağının yapabildiği gibi işlevleri hesaplayabilir. Örneğin, yapay sinir ağları, görüntünün çeşitli insanları tasvir etme olasılığını temsil eden farklı görüntü piksellerinin parlaklığını temsil eden sayıları ve çıktı sayılarını girmek üzere eğitilmiştir. Burada her bir yapay nöron (daire), yukarıdan bağlantılar (hatlar) yoluyla kendisine gönderilen sayıların ağırlıklı toplamını hesaplar, basit bir işlev uygular ve sonucu aşağı doğru iletir, sonraki her katman daha yüksek seviyeli özellikleri hesaplar. Tipik yüz tanıma ağları yüz binlerce nöron içerir; şekil, anlaşılması için sadece bir avuç dolusu göstermektedir.

Bu basit yapay sinir ağlarının başarısı, substrat bağımsızlığının bir başka örneğidir: sinir ağları, yapılarının düşük seviyeli nitty-cesur detaylarından görünüşte bağımsız olarak büyük bir hesaplama gücüne sahiptir. Aslında, George Cybenko, Kurt Hornik, Maxwell Stinchcombe ve Halbert White, 1989'da dikkate değer bir şeyi kanıtladılar: bu kadar basit sinir ağları

evrensel hesaplayabildikleri anlamda *hiç* sadece bu sinaps gücü sayılarını buna göre ayarlayarak keyfi olarak doğru bir şekilde işlev görür. Başka bir deyişle, evrim muhtemelen biyolojik nöronlarımızı gerekli olduğu için bu kadar karmaşık hale getirmedi, ancak daha verimli olduğu için - ve insan mühendislerin aksine evrim, basit ve anlaşılması kolay tasarımları ödüllendirmediği için.

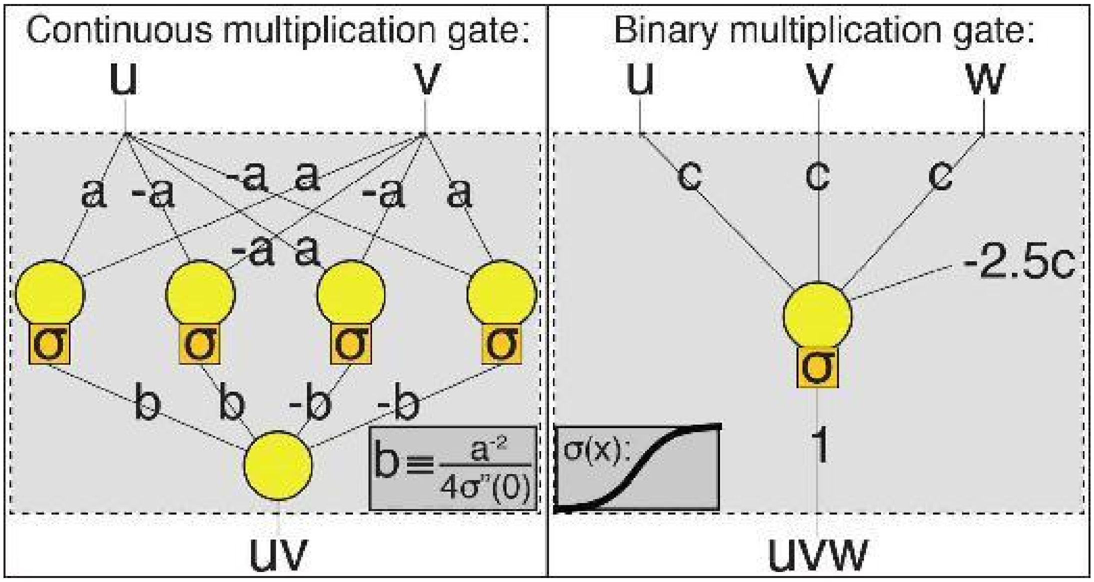
Bunu ilk öğrendiğimde, bu kadar basit bir şeyin rastgele karmaşık bir şeyi nasıl hesaplayabildiğine şaşırmıştım. Örneğin, çarpma kadar basit bir şeyi bile nasıl hesaplayabilirsiniz, her şeye izin verildiğinde

Ağırlıklı toplamları hesaplamak ve tek bir sabit işlev uygulamak mı? Bunun nasıl çalıştığına dair bir tat almak istersen, şekil 2.10 sadece beş nöronun iki rastgele sayıyı nasıl çarptığını ve tek bir nöronun üç biti nasıl çarptığını gösteriyor.

Bununla birlikte, her şeyi hesaplayabileceğinizi kanıtlayabilirsiniz. *teori* keyfi olarak büyük bir sinir ağıyla, kanıt, bunu yapıp yapamayacağınıza dair hiçbir şey söylemiyor. *uygulama,* makul büyüklükte bir ağ ile. Aslında, bunu ne kadar çok düşünürsem, sinir ağlarının bu kadar iyi çalıştığına dair kafam daha da karıştı.

Örneğin, megapiksel gri tonlamalı resimleri iki kategoriye, örneğin kediler veya köpeklere sınıflandırmak istediğimizi varsayalım. Milyon piksellerin her biri bir tane alabilirse 256 değer, sonra 256 değer ^{1000000} olası görüntüler ve her biri için, bir kediyi tasvir etme olasılığını hesaplamak istiyoruz. Bu, bir resim girip bir olasılık çıkaran keyfi bir fonksiyonun bir

256 listesi ^{1000000} Olasılıklar, yani Evrenimizdeki atomlardan çok daha fazla sayı (yaklaşık 10 ^{78}). Yine de, yalnızca binlerce veya milyonlarca parametreye sahip sinir ağları, bir şekilde bu tür sınıflandırma görevlerini oldukça iyi bir şekilde yerine getirmeyi başarır. Bu kadar az parametre gerektirmesi anlamında başarılı sinir ağları nasıl "ucuz" olabilir? Sonuçta, Evrenimize sığacak kadar küçük bir sinir ağının, neredeyse tüm işlevlere yaklaşmakta epeyce başarısız olacağını ve ona atayabileceğiniz tüm hesaplama görevlerinin gülünç derecede küçük bir bölümünü başaracağını kanıtlayabilirsiniz.



Şekil 2.10: Madde nasıl çoğalabilir, ancak aşağıdaki gibi NAND kapıları kullanılmaz şekil 2.7 ama nöronlar. Kilit nokta, ayrıntıları takip etmeyi gerektirmez ve yalnızca nöronların (yapay veya biyolojik) matematik yapabilmesinin yanı sıra çarpma işlemi, NAND kapılarından çok daha az nöron gerektirir. *Zorlu matematik hayranları için isteğe bağlı ayrıntılar:* Daireler toplama yapar, kareler işlevi uygular *σ,* ve çizgiler onları etiketleyen sabitlerle çarpılır. Girişler gerçek sayılar (solda) ve bitlerdir (sağda). Çarpma işlemi keyfi olarak doğru hale gelir. $a \to 0$ (sol) ve $c \to \infty$ (sağ). Sol ağ herhangi bir işlev için çalışır *σ (x)* başlangıçta kavisli (ikinci türevle $\sigma''(0) \neq 0$), Taylor genişlemesi ile kanıtlanabilir *σ (x).* Doğru ağ, işlevin *σ (x)*

0 ve 1'e yaklaştığında *x* sırasıyla çok küçük ve çok büyük olur, bu da $uvw = 1$ yalnızca $u + v + w = 3$. (Bu örnekler, öğrencilerim Henry Lin ve David Rolnick ile yazdığım, http://arxiv.org/abs/1608.08225 adresinde bulunan "Derin ve Ucuz Öğrenme Neden Bu Kadar İyi Çalışıyor?" Makalesinden alınmıştır.) Çok sayıda çarpımı (yukarıdaki gibi) ve toplamayı bir araya getirerek, herhangi bir düzgün işlevi yaklaşık olarak tahmin edebildiği bilinen herhangi bir polinomu hesaplayabilirsiniz.

Öğrencim Henry Lin ile bu ve ilgili gizemler üzerine kafa yorarken çok eğlendim. Hayatta en çok minnettar hissettiğim şeylerden biri harika öğrencilerle işbirliği yapma fırsatı ve Henry de onlardan biri. Onunla çalışmakla ilgilenip ilgilenmediğimi sormak için ofisime ilk geldiğinde, benimle çalışmakla ilgilenip ilgilenmediğini sormanın benim için daha uygun olacağını düşündüm: bu mütevazı, arkadaş canlısı ve parlak gözlü çocuk Louisiana, Shreveport'tan sekiz bilimsel makale yazmış, bir Forbes 30-Under-30 ödülü kazanmış ve bir milyondan fazla izlenme ile bir TED konuşması yapmıştı - ve o sadece yirmi yaşındaydı! Bir yıl sonra şaşırtıcı bir şekilde birlikte bir makale yazdık.

Sonuç: Sinir ağlarının neden bu kadar iyi çalıştığı sorusu tek başına matematikle cevaplanamaz çünkü cevabın bir kısmı fizikte yatıyor. Fizik yasalarının bize attığı ve bizi bilgisayarla ilgilenmeye iten işlevler sınıfının da oldukça küçük bir sınıf olduğunu gördük çünkü hala tam olarak anlamadığımız nedenlerden dolayı, fizik yasaları oldukça basit. Dahası, sinir ağlarının hesaplayabildiği küçük fonksiyon fraksiyonu, fiziğin bizi ilgilendirdiği küçük fraksiyona çok benziyor! Ayrıca, derin öğrenme sinir ağlarının (çok sayıda katman içeriyorlarsa "derin" olarak adlandırılırlar) bu ilgili işlevlerin çoğu için sığ olanlardan çok daha verimli olduğunu gösteren önceki çalışmayı da genişlettik. Örneğin, başka bir harika MIT öğrencisi olan David Rolnick ile birlikte,

çarpma *n* sayılar devasa bir 2 gerektirir n tek katmanlı bir ağ için nöronlar, ancak yalnızca yaklaşık 4 *n* derin bir ağdaki nöronlar. Bu, yapay zeka araştırmacıları arasında neden sinir ağlarının artık tüm öfke olduğunu açıklamaya yardımcı olmakla kalmıyor, aynı zamanda neden beyinlerimizde sinir ağları geliştirdiğimizi de açıklamaya yardımcı oluyor: Geleceği tahmin etmek için beyinleri geliştirmiş olsaydık, o zaman hesaplamalı bir mimari geliştirmiş oluruz. fiziksel dünyada önemli olan hesaplama problemlerinde iyi.

Artık sinir ağlarının nasıl çalıştığını ve hesaplandığını keşfettiğimize göre, nasıl öğrenebilecekleri sorusuna dönelim. Spesifik olarak, bir sinir ağı, sinapslarını güncelleyerek bilgi işlemde nasıl daha iyi hale gelebilir?

1949'daki ufuk açıcı kitabında, *Davranışın Organizasyonu: Nöropsikolojik Bir Teori,* Kanadalı psikolog Donald Hebb, yakınlardaki iki nöronun aynı anda sıklıkla aktif olması ("ateşleme") durumunda, sinaptik eşleşmelerinin güçleneceğini ve böylece birbirlerini tetiklemeye yardımcı olmayı öğreneceklerini savundu - popüler sloganı "Birlikte ateş edin," birlikte bağlayın. " Gerçek beyinlerin nasıl öğrendiğinin ayrıntıları hala anlaşılamamış olsa da ve araştırmalar çoğu durumda cevapların çok daha karmaşık olduğunu göstermiş olsa da, bu basit öğrenme kuralının (Hebbian öğrenme olarak bilinir) bile sinir ağlarının öğrenmesine izin verdiği gösterilmiştir. ilginç şeyler. John Hopfield, Hebbian öğreniminin, aşırı basitleştirilmiş yapay sinir ağının, sadece tekrar tekrar maruz kalarak birçok karmaşık anıyı depolamasına izin verdiğini gösterdi. Bilgiye bu tür bir şekilde maruz kalmak, yapay sinir ağlarına (veya becerilerin öğretilen hayvanlara veya insanlara) atıfta bulunulduğunda genellikle "eğitim" olarak adlandırılır, ancak "çalışmak", "eğitim" veya "deneyim" aynı derecede uygun olabilir. Günümüzün yapay zeka sistemlerine güç veren yapay sinir ağları, Hebbian öğrenimini daha sofistike öğrenme kurallarıyla değiştirmeye meyillidir.

“geri yayılım” ve “stokastik gradyan iniş” gibi isimler, ancak temel fikir aynı: sinapsların zaman içinde güncellendiği bir fizik yasasına benzer basit bir deterministik kural var. Sihir yoluyla, bu basit kural, eğitim büyük miktarda veriyle gerçekleştirilirse sinir ağının dikkat çekici derecede karmaşık hesaplamaları öğrenmesini sağlayabilir. Henüz beynimizin hangi öğrenme kurallarını kullandığını tam olarak bilmiyoruz, ancak cevap ne olursa olsun, fizik yasalarını ihlal ettiklerine dair hiçbir gösterge yok.

Çoğu dijital bilgisayarın çalışmalarını birden çok adıma bölerek ve hesaplama modüllerini defalarca yeniden kullanarak verimlilik elde etmesi gibi, birçok yapay ve biyolojik sinir ağları da öyle. Beyinlerin, bilgisayar bilimcilerinin dediği parçalar vardır *tekrarlayan* Bilginin tek bir yol yerine birden çok yönde akabildiği ileri beslemeli sinir ağlarından ziyade, mevcut çıktının daha sonra olacaklara girdi haline gelebilmesi. Bir dizüstü bilgisayarın mikroişlemcisindeki mantık kapıları ağı da bu anlamda yineleniyor: geçmiş bilgilerini yeniden kullanmaya devam ediyor ve klavyeden, izleme dörtgeninden, kameradan vb. Yeni bilgi girişinin devam eden hesaplamayı etkilemesine izin veriyor ve bu da sırayla örneğin bir ekrana, hoparlöre, yazıcıya veya kablosuz ağa bilgi çıkışı. Benzer şekilde, beyninizdeki nöron ağı tekrarlanır ve gözleriniz, kulaklarınız ve diğer duyulardan bilgi girişinin devam eden hesaplamayı etkilemesine izin verir ve bu da kaslarınıza bilgi çıkışını belirler.

Öğrenmenin tarihi, en azından yaşamın tarihinin kendisi kadardır, çünkü her kendini üreten organizma bilgiyi ilginç bir şekilde kopyalayıp işlemektedir - bir şekilde öğrenilmiş davranış.
Yaşam döneminde
Ancak 1.0, organizmalar yaşamları boyunca öğrenmediler: bilgiyi işleme ve tepki verme kuralları, kalıtsal DNA'ları tarafından belirlendi, bu nedenle tek öğrenme, nesiller boyunca Darwinci evrim yoluyla tür düzeyinde yavaş bir şekilde gerçekleşti.

Yaklaşık yarım milyar yıl önce, Dünya'daki bazı gen hatları, sinir ağları içeren hayvanları yaşam boyunca deneyimlerden davranışları öğrenebilecek hale getirmenin bir yolunu keşfetti. Life 2.0 geldi ve dramatik bir şekilde daha hızlı öğrenme ve rekabeti alt etme yeteneği nedeniyle, tüm dünyaya orman yangını gibi yayıldı. Bölüm 1'de incelediğimiz gibi, yaşam öğrenmede giderek daha iyi hale geldi ve gittikçe artan bir hızda. Maymun benzeri belirli bir tür, bilgi edinme konusunda o kadar becerikli bir beyin geliştirdi ki, alet kullanmayı, ateş yakmayı, dil konuşmayı ve karmaşık bir küresel toplum yaratmayı öğrendi. Bu toplumun kendisi, bir icat bir sonrakini mümkün kılarken, hepsini hızlandıran bir hızla hatırlayan, hesaplayan ve öğrenen bir sistem olarak görülebilir: yazı, matbaa, modern

bilim, bilgisayarlar, internet vb. Gelecekteki tarihçiler, icatları olanaklı kılan bu listeye bundan sonra ne koyacaklar? Benim tahminim yapay zeka.

Hepimizin bildiği gibi, bilgisayar belleğindeki ve hesaplama gücündeki büyük gelişmeler (şekil 2.4 ve şekil 2.8) yapay zeka alanında olağanüstü bir ilerlemeye dönüştü - ancak makinenin *öğrenme*

yaş geldi. IBM'in Deep Blue bilgisayarı 1997'de satranç şampiyonu Garry Kasparov'u alt ettiğinde, başlıca avantajları öğrenmede değil bellek ve hesaplamada yatıyordu. Hesaplamalı zekası, bir insan ekibi tarafından yaratılmıştı ve Deep Blue'nun yaratıcılarını geride bırakmasının temel nedeni, daha hızlı hesaplama ve dolayısıyla daha fazla potansiyel pozisyonu analiz etme yeteneğiydi. IBM'in Watson bilgisayarı, yarışma programında insan dünya şampiyonunu tahttan indirdiğinde

Jeopardy !, özel programlanmış becerilere ve üstün bellek ve hıza göre öğrenmeye daha az dayanıyordu. Bacaklı hareketten kendi kendine giden arabalara ve kendi kendine inen roketlere kadar robotikteki çoğu erken atılım için de aynı şey söylenebilir.

Buna karşılık, en son yapay zeka buluşlarının arkasındaki itici güç, makine *öğrenme.* Düşünmek şekil 2.11 , Örneğin. Neyin fotoğrafı olduğunu söylemeniz kolaydır, ancak bir görüntünün tüm piksellerinin renklerinden başka hiçbir şey girmeyen ve "Frizbi oyunu oynayan bir grup genç" gibi doğru bir başlık veren bir işlevi programlamak onlarca yıldır dünyadaki tüm yapay zeka araştırmacılarından kaçtı. Yine de, Ilya Sutskever liderliğindeki bir Google ekibi, 2014 yılında tam olarak bunu yaptı. Farklı bir piksel renkleri seti girdiğinizde, "Kuru çim tarlada yürüyen bir fil sürüsü" yine doğru bir şekilde yanıt veriyor. Bunu nasıl yaptılar? Frizbi, yüz ve benzerlerini algılamak için el yapımı algoritmalar programlayarak Deep Blue tarzı? Hayır, fiziksel dünya veya içeriği hakkında hiçbir bilgisi olmayan nispeten basit bir sinir ağı oluşturarak ve sonra onu büyük miktarda veriye maruz bırakarak öğrenmesine izin vererek.



Şekil 2.11: "Bir frizbi oyunu oynayan bir grup genç" - bu başlık, insanları, oyunları veya frizbi anlamayan bir bilgisayar tarafından yazılmıştır.

Çocuklarımızın nasıl öğrendiğini tam olarak anlamadığımız gibi, bu tür sinir ağlarının nasıl öğrendiğini ve neden ara sıra başarısız olduklarını tam olarak anlamıyoruz. Ancak açık olan şey, halihazırda oldukça yararlı oldukları ve derin öğrenmeye yönelik bir yatırım dalgasını tetikledikleri. Derin öğrenme, artık el yazısı transkripsiyonundan sürücüsüz otomobiller için gerçek zamanlı video analizine kadar bilgisayarla görmenin birçok yönünü değiştirdi. Benzer şekilde, bilgisayarların konuşulan dili metne dönüştürme ve gerçek zamanlı olarak bile başka dillere çevirme becerisinde devrim yarattı - bu yüzden artık Siri, Google Now ve Cortana gibi kişisel dijital asistanlarla konuşabiliyoruz. Bir web sitesini insan olduğumuza ikna etmemiz gereken sinir bozucu CAPTCHA bulmacaları, makine öğrenimi teknolojisinin yapabileceklerinin önüne geçmek için gittikçe zorlaşıyor. 2015 yılında Google DeepMind, bir çocuk gibi düzinelerce bilgisayar oyununda ustalaşabilen derin öğrenmeyi kullanan bir yapay zeka sistemini piyasaya sürdü - hiçbir talimat olmadan - kısa süre sonra herhangi bir insandan daha iyi oynamayı öğrendi. 2016 yılında aynı şirket, farklı tahta konumlarının gücünü değerlendirmek için derin öğrenmeyi kullanan ve dünyanın en güçlü Go şampiyonunu mağlup eden Go-oynayan bir bilgisayar sistemi olan AlphaGo'yu kurdu. Bu ilerleme erdemli bir çemberi besliyor, Farklı tahta konumlarının gücünü değerlendirmek için derin öğrenmeyi kullanan ve dünyanın en güçlü Go şampiyonunu mağlup eden bir Go-oynayan bilgisayar sistemi. Bu ilerleme erdemli bir çemberi besliyor, Farklı tahta konumlarının gücünü değerlendirmek için derin öğrenmeyi kullanan ve dünyanın en güçlü Go şampiyonunu mağlup eden bir Go-oynayan bilgisayar sistemi. Bu ilerleme erdemli bir çemberi besliyor,

Yapay zeka araştırmalarına daha fazla fon ve yetenek getirerek daha fazla ilerleme sağlar.

Bu bölümü şimdiye kadar zekanın doğasını ve gelişimini araştırarak geçirdik. Makinelerin bizim için rekabete girmesi ne kadar sürer? *herşey* bilişsel görevler? Açıkça bilmiyoruz ve cevabın "asla" olabileceği olasılığına açık olmamız gerekiyor. Bununla birlikte, bu bölümün temel bir mesajı şudur ki, bunun olasılığını da göz önünde bulundurmamız gerekir. *niyet* Olabilir, belki bizim hayatımızda bile. Sonuçta, madde fizik kanunlarına uyduğunda hatırlayacak, hesaplayacak ve öğrenecek ve maddenin biyolojik olması gerekmeyecek şekilde düzenlenebilir. Yapay zeka araştırmacıları genellikle fazla vaatte bulunmakla ve yetersiz hizmet vermekle suçlanıyor, ancak adaletli olmak gerekirse, eleştirmenlerinden bazıları en iyi performansa sahip değil. Bazıları hedef direklerini hareket ettirmeye devam ediyor, zekayı bilgisayarların hala yapamadığı veya bizi etkileyen şey olarak tanımlıyor. Makineler artık aritmetik, satranç, matematiksel teorem kanıtlama, hisse senedi toplama, resim yazısı oluşturma, sürüş, atari oyunu oynama, Go, konuşma sentezi, konuşma transkripsiyonu, çeviri ve kanser teşhisinde iyi veya mükemmel, ancak bazı eleştirmenler küçümseyici bir şekilde "Tabii ki" alay edecekler. bu değil *gerçek*

zeka!" Gerçek zekanın yalnızca Moravec'in arazisindeki dağ zirvelerini içerdiğini iddia edebilirler (şekil 2.2), tıpkı geçmişte bazı insanların, su yükselmeye devam ederken resim yazısı ve Go'nun sayılması gerektiğini savunduğu gibi, henüz su altında kalmamış.

Suyun en az bir süre daha yükselmeye devam edeceğini varsayarsak, AI'nın toplum üzerindeki etkisi artmaya devam edecek. Yapay zeka, tüm görevlerde insan seviyesine ulaşmadan çok önce, bize böcekler, kanunlar, silahlar ve işler gibi konuları içeren büyüleyici fırsatlar ve zorluklar sağlayacaktır. Bunlar nedir ve onlar için en iyi nasıl hazırlanabiliriz? Bunu bir sonraki bölümde inceleyelim.

ALT ÇİZGİ:

- Karmaşık hedefleri gerçekleştirme yeteneği olarak tanımlanan zeka, tek bir IQ ile ölçülemez, sadece tüm hedeflerdeki bir yetenek spektrumuyla ölçülebilir.

- Günümüzün yapay zekası, *dar,* her sistem yalnızca çok özel hedefleri gerçekleştirebilirken, insan zekası dikkate değer şekilde *kalın.*

- Hafıza, hesaplama, öğrenme ve zeka onlara soyut, soyut ve ruhani bir his veriyor çünkü *substrattan bağımsız:* Altta yatan malzeme alt tabakasının ayrıntılarına bağlı olmayan veya bunları yansıtmayan kendi hayatlarını sürdürebilirler.

- Herhangi bir madde parçası, *hafıza* birçok farklı kararlı duruma sahip olduğu sürece.

- Herhangi bir konu olabilir *computronium,* için substrat *hesaplama,* herhangi bir işlevi uygulamak için birleştirilebilen belirli evrensel yapı bloklarını içerdiği sürece. NAND geçitleri ve nöronlar, bu tür evrensel "hesaplamalı atomların" iki önemli örneğidir.

- Sinir ağı, aşağıdakiler için güçlü bir alt tabakadır: *öğrenme* çünkü, sadece fizik kanunlarına uyarak, istenen hesaplamaları daha iyi ve daha iyi hale getirmek için kendini yeniden düzenleyebilir.

- Fizik kanunlarının çarpıcı basitliğinden dolayı, biz insanlar akla gelebilecek tüm hesaplama problemlerinin sadece küçük bir kısmını önemsiyoruz ve sinir ağları tam da bu küçük fraksiyonu çözmede dikkat çekici derecede iyi olma eğilimindeyiz.

- Teknoloji iki kat daha güçlü hale geldiğinde, genellikle iki kat daha güçlü bir teknoloji tasarlamak ve inşa etmek için kullanılabilir ve Moore yasasının ruhunda tekrarlanan kapasitenin ikiye katlanmasını tetikler. Bilgi teknolojisinin maliyeti şu anda yaklaşık bir yüzyıl boyunca her iki yılda bir yarı yarıya azaldı ve bilgi çağını mümkün kıldı.

- AI ilerlemesi devam ederse, AI tüm beceriler için insan seviyesine ulaşmadan çok önce, bize böcekler, kanunlar, silahlar ve işler gibi konuları içeren - bir sonraki bölümde inceleyeceğimiz - büyüleyici fırsatlar ve zorluklar sağlayacaktır.

---

* 1 Bunu görmek için, birisi Olimpik düzeyde atletik başarılar elde etme yeteneğinin "atletik bölüm" veya kısaca AQ olarak adlandırılan tek bir sayı ile ölçülebileceğini iddia ederse nasıl tepki vereceğinizi hayal edin, böylece en yüksek AQ'ya sahip Olimpiyat tüm spor dallarında altın madalya kazanacaktı.

* 2 Bazı insanlar AGI ile eşanlamlı olarak "insan düzeyinde AI" veya "güçlü AI" yı tercih ederler, ancak her ikisi de sorunludur. Bir cep hesap makinesi bile dar anlamda insan düzeyinde bir AI'dır. "Güçlü yapay zeka" nın zıtlığı "zayıf yapay zeka" dır, ancak Deep Blue, Watson ve AlphaGo gibi dar yapay zeka sistemlerini "zayıf" olarak adlandırmak garip geliyor.

* 3 NAND, NOT AND'ın kısaltmasıdır: Bir AND geçidi, yalnızca ilk giriş 1 ise ve ikinci giriş ise 1 çıktısı verir. 1, dolayısıyla NAND tam tersini verir.

* 4 Matematikçilerin ve bilgisayar bilimcilerinin "hesaplanabilir işlev" dedikleri, yani sınırsız bellek ve zamana sahip bir varsayımsal bilgisayar tarafından hesaplanabilen bir işlevi kastetmek için "iyi tanımlanmış işlevi" kullanıyorum. Alan Turing ve Alonzo Kilisesi, tanımlanabilen ancak hesaplanamayan işlevlerin de olduğunu ünlü bir şekilde kanıtladı.

* 5 Matematiği seviyorsanız, bu etkinleştirme işlevinin iki popüler seçeneği sözde sigmoid işlevidir.

$\sigma(x) \equiv 1/(1+e^{-x})$ ve rampa fonksiyonu $\sigma(x) =$ en fazla $\{0, x\}$, Doğrusal olmadığı sürece (düz bir çizgi) hemen hemen her işlevin yeterli olacağı kanıtlanmış olmasına rağmen. Hopfield'ın ünlü modeli $\sigma(x) = -1$ eğer $x < 0$ ve $\sigma(x) = 1$ eğer $x \geq 0$. Eğer nöron durumları bir vektörde saklanırsa, ağ, bu vektörü basitçe sinaptik bağlaşımları depolayan bir matrisle çarparak ve ardından tüm elemanlara σ fonksiyonunu uygulayarak güncellenir.

Bölüm 3

Yakın Gelecek: Atılımlar, Hatalar, Kanunlar, Silahlar ve İşler

Yakında yön değiştirmezsek, gittiğimiz yere varırız.

Irwin Corey

Günümüzde insan olmak ne anlama geliyor? Örneğin, kendimiz hakkında gerçekten değer verdiğimiz, bizi diğer yaşam formlarından ve makinelerden farklı kılan nedir? Başkaları bizim hakkımızda, bazılarını bize iş teklif etmeye istekli kılan neye değer veriyor? Herhangi bir zamanda bu sorulara verdiğimiz yanıtlar ne olursa olsun, teknolojinin yükselişinin onları yavaş yavaş değiştirmesi gerektiği açıktır.

Mesela beni al. Bir bilim insanı olarak, kendi hedeflerimi belirlemekten, çok çeşitli çözülmemiş sorunların üstesinden gelmek için yaratıcılığı ve sezgiyi kullanmaktan ve keşfettiklerimi paylaşmak için dili kullanmaktan gurur duyuyorum. Neyse ki benim için toplum bunu bir iş olarak yapmam için bana para vermeye razı. Yüzyıllar önce, ben de diğerleri gibi ben de kimliğimi bir çiftçi veya zanaatkar olma üzerine inşa edebilirdim, ancak teknolojinin büyümesi o zamandan beri bu tür meslekleri işgücünün çok küçük bir kısmına indirdi. Bu, artık herkesin kendi kimliğini çiftçilik veya el sanatları etrafında inşa etmesinin mümkün olmadığı anlamına geliyor.

Şahsen, bugünün makinelerinin kazma ve örgü gibi el becerilerinde beni geride bırakması beni rahatsız etmiyor, çünkü bunlar ne hobilerim ne de gelir kaynaklarım veya öz değerim. Nitekim, bu konudaki yeteneklerimle ilgili sahip olabileceğim herhangi bir kuruntu, okulum beni neredeyse başarısız olduğum bir örgü dersi almaya zorladığında ve okulumu bitirdiğimde sekiz yaşında ezildi.

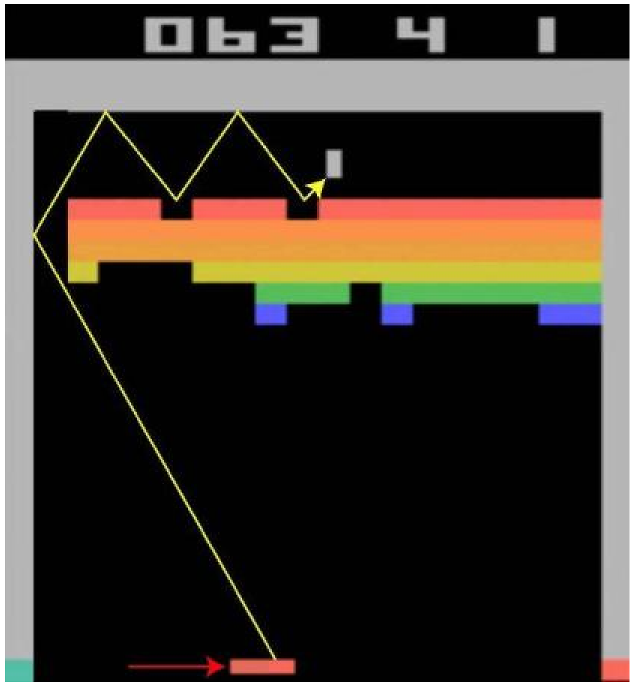
sadece bana acıyan beşinci sınıftan şefkatli bir yardımcı sayesinde proje.

Ancak teknoloji gelişmeye devam ettikçe, yapay zekanın yükselişi, iş piyasasında şu anki öz değer ve değer duygumu sağlayan yetenekleri de gölgede bırakacak mı? Stuart Russell bana kendisinin ve diğer yapay zeka araştırmacılarının birçoğunun yakın zamanda "kutsal bir bok" yaşadığını söyledi. YZ'nin yıllardır görmeyi beklemedikleri bir şeyi yaptığına şahit olduklarında. Bu ruhla, lütfen size kendi HS anlarımdan birkaçını ve onları yakında geçilecek insan yeteneklerinin habercileri olarak nasıl gördüğümü anlatmama izin verin.

Buluşlar

Deep Reinforcement Learning Agents

Bilgisayar oyunları oynamayı öğrenen bir DeepMind AI sisteminin videosunu izlerken 2014 yılında en büyük çene düşüşlerimden birini yaşadım. Özellikle, AI Breakout oynuyordu (bkz. şekil 3.1), gençlerimden sevgiyle hatırladığım klasik bir Atari oyunu. Amaç, bir topu tuğla duvardan defalarca sektirmek için bir raketle manevra yapmaktır; Bir tuğlaya her vurduğunuzda, kaybolur ve puanınız artar.



Şekil 3.1: Atari oyunu Breakout'u sıfırdan oynamayı öğrendikten sonra, skoru en üst düzeye çıkarmak için derin pekiştirmeli öğrenmeyi kullanarak, DeepMind AI en uygun stratejiyi keşfetti: tuğla duvarın en sol kısmında bir delik açmak ve topun arkada zıplamasına izin vermek çok hızlı bir şekilde puan topluyor. Top ve raketin geçmiş yörüngelerini gösteren oklar çizdim.

O gün kendime ait bazı bilgisayar oyunları yazmıştım ve Breakout'u oynayabilecek bir program yazmanın zor olmadığının farkındaydım - ama DeepMind ekibinin yaptığı şey bu değildi. Bunun yerine, bu oyun hakkında veya diğer oyunlar hakkında ve hatta hakkında hiçbir şey bilmeyen boş bir sayfa yapay zekası yarattılar.
kavramlar oyunlar, kürekler, tuğlalar veya toplar gibi. Yapay zekalarının bildiği tek şey, düzenli aralıklarla uzun bir sayı listesinin beslendiğiydi: mevcut skor ve ekranın farklı bölümlerinin nasıl renklendirildiğinin özellikleri olarak bizim (ancak YZ'nin değil) tanıyacağımız uzun bir sayı listesi. Yapay zekaya, belirli aralıklarla, bizim (ancak YZ'nin değil) tuşlara basacağımız kodlar olarak tanıyacağımız sayıları çıkararak puanı maksimize etmesi söylendi.

Başlangıçta, AI çok kötü oynadı: Görünüşe göre rastgele bir şekilde raketi ileri geri salladı ve neredeyse her seferinde topu ıskaladı. Bir süre sonra, küreği topa doğru hareket ettirmenin iyi bir fikir olduğu anlaşıldı.

Fikir, çoğu zaman hala ıskalamasına rağmen. Ancak pratikle gelişmeye devam etti ve kısa süre sonra oyunda hiç olmadığım kadar iyi oldu, ne kadar hızlı yaklaşırsa yaklaşsın topu yanılmadan geri getirdi. Ve sonra çenem düştü: her zaman sol üst köşenin duvardan bir delik açmasını ve topun duvarın arkası ile arkasındaki bariyer arasında zıplamasına izin vermesini hedefleyen bu şaşırtıcı skor maksimize etme stratejisini buldu. Bu gerçekten akıllıca bir şey gibi geldi. Nitekim Demis Hassabis daha sonra bana DeepMind ekibindeki programcıların, inşa ettikleri yapay zekadan öğrenene kadar bu hileyi bilmediklerini söyledi. Adresinde kendinize bunun bir videosunu izlemenizi tavsiye ederim

sağladığım bağlantı. [1]

Bunda biraz rahatsız edici bulduğum insan benzeri bir özellik vardı: Bir hedefi olan ve ona ulaşmada her zamankinden daha iyi olmayı öğrenen ve sonunda yaratıcılarından daha iyi performans gösteren bir YZ izliyordum. Önceki bölümde, zekayı basitçe karmaşık hedeflere ulaşma yeteneği olarak tanımlamıştık, bu nedenle, DeepMind'in yapay zekası gözlerimin önünde daha akıllı hale geliyordu (her ne kadar bu belirli oyunu oynamanın çok dar anlamında olsa da). İlk bölümde bilgisayar bilimcilerin dedikleri şeyle karşılaştık *akıllı ajanlar:* algılayıcılardan çevreleri hakkında bilgi toplayan ve daha sonra bu bilgileri işleyerek çevrelerine nasıl tepki vereceklerine karar veren varlıklar. DeepMind'ın oyun oynayan yapay zekası tuğlalar, kürekler ve toplardan oluşan son derece basit bir sanal dünyada yaşamasına rağmen, bunun akıllı bir ajan olduğunu inkar edemedim.

DeepMind çok geçmeden yöntemini yayınladı ve kodunu paylaşarak, çok basit ama güçlü bir fikir olduğunu açıkladı. *derin pekiştirmeli öğrenme.* [2]

Temel pekiştirmeli öğrenme, davranışçı psikolojiden ilham alan klasik bir makine öğrenimi tekniğidir; burada olumlu bir ödül almak, bir şeyi tekrar yapma eğiliminizi artırır ve bunun tersi de geçerlidir. Tıpkı bir köpeğin, yakında sahibinden cesaret alma veya atıştırmalık alma olasılığını artırdığında hileler yapmayı öğrenmesi gibi, DeepMind'in yapay zekası topu yakalamak için raketi hareket ettirmeyi öğrendi çünkü bu, yakında daha fazla puan alma olasılığını artırdı. DeepMind bu fikri derin öğrenmeyle birleştirdi: klavyede izin verilen tuşların her birine basarak ortalama olarak kaç puan kazanılacağını tahmin etmek için bir önceki bölümde olduğu gibi derin bir sinir ağı eğittiler ve ardından AI, hangi tuşa basarsa onu seçti. Oyunun mevcut durumuna bakıldığında sinir ağı en umut verici olarak değerlendirildi.

Kişisel öz-değer duyguma katkıda bulunan özellikleri listelediğimde,

insan, geniş bir yelpazede çözülmemiş problemlerin üstesinden gelme yeteneğini dahil ettim. Aksine, Breakout'u oynayabilmek ve başka hiçbir şey yapmamak son derece dar bir zeka oluşturur. Bana göre, DeepMind'in atılımının gerçek önemi, derin pekiştirmeli öğrenmenin tamamen genel bir teknik olmasıdır. Elbette, aynı yapay zeka uygulamasının kırk dokuz farklı Atari oyunu oynamasına izin verdiler ve Pong'dan Boks'a, Video Pinball'dan Uzay İstilacılarına kadar yirmi dokuzunda insan test oyuncularını geride bırakmayı öğrendiler.

Aynı AI fikrinin, dünyaları iki boyutlu yerine üç boyutlu olan daha modern oyunlarda kendini kanıtlamaya başlaması uzun sürmedi. Yakında DeepMind'in OpenAI'deki San Francisco merkezli rakipleri, DeepMind'in yapay zekasının ve diğer akıllı ajanlarının tüm bilgisayarla bir oyunmuş gibi etkileşimde bulunabileceği bir platform yayınladı: herhangi bir şeye tıklamak, herhangi bir şeyi yazmak ve her ne yazılımsa onu açmak ve çalıştırmak. Örneğin, bir web tarayıcısını çalıştırıp çevrimiçi ortamda gezinmek gibi.

Derin pekiştirmeli öğrenmenin ve bunun üzerine yapılan iyileştirmelerin geleceğine baktığımızda, görünürde açık bir son yok. Potansiyel sanal oyun dünyalarıyla sınırlı değildir, çünkü bir robotsanız, hayatın kendisi bir oyun olarak görülebilir. Stuart Russell bana, ilk büyük HS anının, Big Dog'un karla kaplı bir orman yamacında koşarken, bacaklı hareketi zarif bir şekilde çözmesini izlemek olduğunu söyledi.

Kendisinin yıllarca çözmek için uğraştığı sorunu. [3] Yine de bu dönüm noktasına 2008'de ulaşıldığında, zeki programcıların büyük miktarda çalışmasını gerektiriyordu. DeepMind'ın atılımından sonra, bir robotun insan programcıların yardımı olmadan yürümeyi öğretmek için nihayetinde derin pekiştirmeli öğrenmenin bazı varyantlarını kullanamaması için hiçbir neden yok: gereken tek şey, ilerleme kaydettiğinde ona puan veren bir sistemdir. Gerçek dünyadaki robotlar benzer şekilde, insan programcıların yardımı olmadan yüzmeyi, uçmayı, masa tenisi oynamayı, dövüşmeyi ve diğer motor görevlerinin neredeyse sonsuz bir listesini gerçekleştirmeyi öğrenme potansiyeline sahiptir. İşleri hızlandırmak ve öğrenme sürecinde takılıp kalma veya kendilerine zarar verme riskini azaltmak için, muhtemelen öğrenmelerinin ilk aşamalarını sanal gerçeklikte yapacaklardır.

Sezgi, Yaratıcılık ve Strateji

Benim için bir başka belirleyici an, DeepMind AI sistemi AlphaGo'nun, genellikle 21. yüzyılın başlarında dünyanın en iyi oyuncusu olarak kabul edilen Lee Sedol'e karşı beş maçlık bir Go maçı kazandığı zamandı.

İnsan Go oyuncularının, yirmi yıl önce satranç oynayan meslektaşlarının başına geldiği için, bir noktada makineler tarafından tahttan indirilmesi yaygın bir şekilde bekleniyordu. Bununla birlikte, çoğu Go uzmanı bunun bir on yıl daha süreceğini öngördü, bu yüzden AlphaGo'nun zaferi hem onlar hem de benim için çok önemli bir andı. Nick Bostrom ve Ray Kurzweil, ilk üç maçı kaybetmeden önce ve sonra Lee Sedol ile yaptığı röportajlardan da anlaşılacağı üzere AI atılımlarının geldiğini görmenin ne kadar zor olabileceğini vurguladılar:

- Ekim 2015: "Görülen seviyesine göre ... Oyunu neredeyse heyelanla kazanacağımı düşünüyorum."
- Şubat 2016: "Google DeepMind'in yapay zekasının şaşırtıcı derecede güçlü olduğunu ve daha da güçlendiğini duydum, ancak en azından bu sefer kazanabileceğime eminim."
- 9 Mart 2016: "Kaybedeceğimi düşünmediğim için çok şaşırdım."
- 10 Mart 2016: “Oldukça sessizim… Şoktayım. Kabul edebilirim ki ... üçüncü oyun benim için kolay olmayacak. "
- 12 Mart 2016: "Kendimi biraz güçsüz hissettim."

Lee Sedol oynadıktan sonraki bir yıl içinde, daha da geliştirilmiş bir AlphaGo, tek bir maç kaybetmeden dünyanın en iyi yirmi oyuncusunu oynadı.

Kişisel olarak bu neden benim için bu kadar önemliydi? Yukarıda, sezgi ve yaratıcılığı temel insan özelliklerimden ikisi olarak gördüğümü itiraf ettim ve şimdi açıklayacağım gibi, AlphaGo'nun her ikisini de gösterdiğini hissediyorum.

Go oyuncuları sırayla 19'a 19 tahtaya siyah ve beyaz taşları yerleştirir (bkz. şekil 3.2). Evrenimizdeki atomlardan çok daha fazla olası Go pozisyonu vardır, bu da gelecekteki hareketlerin tüm ilginç dizilerini analiz etmeye çalışmanın hızla umutsuz hale geldiği anlamına gelir. Oyuncular bu nedenle büyük ölçüde

hangi pozisyonların güçlü ve hangilerinin zayıf olduğu neredeyse tekinsiz bir his geliştiren uzmanlarla bilinçli akıl yürütmelerini tamamlayacak bilinçaltı sezgileri. Son bölümde gördüğümüz gibi, derin öğrenmenin sonuçları bazen sezgiyi anımsatır: Derin bir sinir ağı, bir görüntünün nedenini açıklayamadan bir kediyi tasvir ettiğini belirleyebilir. DeepMind ekibi bu nedenle, derin öğrenmenin sadece kedileri değil, aynı zamanda güçlü Go pozisyonlarını da tanıyabileceği fikri üzerine kumar oynadı. AlphaGo'da oluşturdukları temel fikir, derin öğrenmenin sezgisel gücünü GOFAI'nin mantıksal gücü ile birleştirmekti - ki bu, derin öğrenme devriminden önce "Güzel Eski Moda Yapay Zeka" olarak bilinen şeyi ifade ediyor. Hem insan oyunlarından hem de AlphaGo'nun kendi klonunu oynadığı oyunlardan devasa bir Go pozisyonları veritabanı kullandılar. ve her pozisyondan beyazın nihayetinde kazanma olasılığını tahmin etmek için derin bir sinir ağı eğitti. Ayrıca, olası sonraki hamleleri tahmin etmek için ayrı bir ağ eğittiler. Daha sonra bu ağları, yolun aşağısındaki en güçlü konuma götürecek bir sonraki hareketi belirlemek için olası gelecekteki hareket dizilerinin budanmış bir listesini akıllıca araştıran bir GOFAI yöntemiyle birleştirdiler.



Şekil 3.2: DeepMind'in AlphaGo AI'si, 5. satırda, binlerce yıllık insan bilgeliğine meydan okuyarak oldukça yaratıcı bir hamle yaptı ve yaklaşık elli hamle daha sonra Go efsanesi Lee Sedol'u yenmesi için çok önemli olduğunu kanıtladı.

Bu sezgi ve mantığın evliliği, yalnızca güçlü değil, aynı zamanda bazı durumlarda oldukça yaratıcı olan hareketleri de doğurdu. Örneğin, binlerce yıllık Go bilgeliği, oyunun başlarında üçüncü veya dördüncü sırada bir kenardan oynamanın en iyisi olduğunu söylüyor. İkisi arasında bir değiş tokuş var: üçüncü hatta oynamak, tahtanın tarafına doğru kısa vadeli bölge kazanımına yardımcı olurken, dördüncü yardımda oynamak merkeze doğru uzun vadeli stratejik etki sağlar.

İkinci oyunun otuz yedinci hamlesinde AlphaGo, bu kadim bilgeliğe meydan okuyarak ve beşinci hatta oynayarak Go dünyasını şok etti (şekil 3.2), sanki uzun vadeli planlama yeteneklerinde bir insandan bile daha emin ve bu nedenle kısa vadeli kazanç yerine stratejik avantajı tercih ediyormuş gibi. Yorumcular sersemlemişti ve Lee Sedol ayağa kalktı ve geçici olarak odadan çıktı. [4] Yeterince kesin, yaklaşık elli hamle sonra, tahtanın sol alt köşesinden gelen kavga, otuz yedinci hamleden o siyah taşla birleşerek sona erdi! Ve bu motif, nihayetinde oyunu kazanan, AlphaGo'nun beşinci sıradaki hamlesinin mirasını Go tarihinin en yaratıcılarından biri olarak sağlamlaştıran şeydi.

Sezgisel ve yaratıcı yönleri nedeniyle Go, başka bir oyundan çok bir sanat formu olarak görülüyor. Resim, kaligrafi ve kaligrafi ile birlikte antik Çin'de dört "temel sanat" tan biri olarak kabul edildi. *qin* müzik ve AlphaGo ve Lee Sedol arasındaki ilk oyunu neredeyse 300 milyon kişinin izlediği Asya'da oldukça popüler olmaya devam ediyor. Sonuç olarak, Go dünyası sonuçtan oldukça sarsıldı ve AlphaGo'nun zaferini,

insanlık. O dönemde dünyanın en iyi Go oyuncusu Ke Jie şunları söyledi: [5]
“İnsanlık binlerce yıldır Go oynadı ve yine de yapay zekanın bize gösterdiği gibi, henüz yüzeyi çizmedik bile… İnsan ve bilgisayar oyuncularının birliği yeni bir çağ başlatacak… İnsan ve yapay zeka birlikte Go gerçeği. " Bu tür verimli insan-makine işbirliği gerçekten de, yapay zekanın bize insanların anlayışımızı derinleştirmesine ve nihai potansiyelimizi gerçekleştirmesine yardım edebileceğini umduğumuz bilim de dahil olmak üzere birçok alanda umut verici görünüyor.

Bana göre AlphaGo, yakın gelecek için bize başka önemli bir ders de öğretiyor: derin öğrenme sezgisini GOFAI mantığıyla birleştirmek, rakipsiz üretebilir *strateji.* Go nihai strateji oyunlarından biri olduğu için yapay zeka, artık en iyi insan stratejistlerine oyun tahtalarının da ötesinde, örneğin yatırım stratejisi, politik strateji ve askeri strateji ile mezun olmaya ve onlara meydan okumaya (veya yardım etmeye) hazır. Bu tür gerçek dünya strateji sorunları tipik olarak insan psikolojisi, eksik bilgiler ve rastgele olarak modellenmesi gereken faktörler nedeniyle karmaşıktır, ancak poker oynayan AI sistemleri, bu zorlukların hiçbirinin aşılmaz olmadığını zaten göstermiştir.

Doğal lisan

Yapay zeka ilerlemesinin son zamanlarda beni şaşırttığı bir başka alan da dil. Hayatımın erken dönemlerinde seyahate aşık oldum ve diğer kültürler ve diller hakkındaki merak, kimliğimin önemli bir parçasını oluşturdu. İsveççe ve İngilizce konuşarak büyüdüm, okulda Almanca ve İspanyolca öğrettim, iki evlilik yoluyla Portekizce ve Romence öğrendim ve kendime eğlenmek için biraz Rusça, Fransızca ve Mandarin öğrettim.

Ancak AI ulaşıyor ve 2016'daki önemli bir keşfin ardından, Google'ın beyninin ekipmanı tarafından geliştirilen AI sisteminden daha iyi çeviri yapabileceğim neredeyse hiç tembel dil yok.

Kendimi berrak mı yaptım? Aslında şunu söylemeye çalışıyordum:

Ancak AI beni yakalıyor ve 2016'daki büyük bir atılımdan sonra, Google Brain ekibi tarafından geliştirilen AI sisteminden daha iyisi arasında çeviri yapabileceğim neredeyse hiç dil kalmadı.

Ancak, bunu önce İspanyolcaya çevirdim ve birkaç yıl önce dizüstü bilgisayarıma yüklediğim bir uygulamayı kullanarak geri çevirdim. 2016 yılında, Google Brain ekibi ücretsiz Google Translate hizmetini derin ve tekrarlayan sinir ağlarını kullanacak şekilde yükseltti ve eski GOFAI sistemlerine göre gelişme çarpıcıydı: [6]

Ancak AI beni yakalıyor ve 2016'daki bir atılımdan sonra, Google Brain ekibi tarafından geliştirilen AI sisteminden daha iyi çeviri yapabilecek neredeyse hiçbir dil kalmadı.

Gördüğünüz gibi, İspanyolca dolambaçlı yoldan giderken “ben” zamiri kayboldu ve bu maalesef anlamını değiştirdi. Kapat, ama puro yok! Bununla birlikte, Google'ın yapay zekasını savunmak için, ayrıştırılması zor gereksiz uzun cümleler yazdığım için sık sık eleştirilirim ve bu örnek için en kafa karıştırıcı şekilde kıvrımlı olanlardan birini seçtim. Daha tipik cümleler için, yapay zekaları genellikle kusursuz bir şekilde tercüme eder. Sonuç olarak, çıktığında büyük bir heyecan yarattı ve her gün yüz milyonlarca insan tarafından kullanılacak kadar yardımcı oldu. Dahası, konuşmadan metne ve metinden sese dönüştürme için derin öğrenmedeki son gelişmeler sayesinde, bu kullanıcılar artık akıllı telefonlarıyla tek bir dilde konuşabilir ve çevrilen sonucu dinleyebilir.

Doğal dil işleme artık yapay zekanın en hızlı gelişen alanlarından biri ve bence daha fazla başarının büyük bir etkisi olacak çünkü dil

insan olmanın çok merkezinde. Bir YZ, dil tahmininde ne kadar iyi olursa, makul e-posta yanıtları oluşturabilir veya sözlü bir sohbete o kadar iyi devam edebilir. Bu, en azından bir yabancı için, gerçekleşen insan düşüncesi görüntüsünü verebilir. Böylece derin öğrenme sistemleri, bir makinenin bir kişiyi kandırarak kendisinin de insan olduğunu düşünmesi için yazılı olarak yeterince iyi konuşması gereken ünlü Turing testini geçme yolunda küçük adımlar atıyor.

Yine de, dil işleme yapay zekasının gidecek uzun bir yolu var. Bir yapay zeka tarafından çevrildiğimde kendimi biraz sönük hissettiğimi itiraf etmem gerekse de, kendime şunu hatırlattığımda daha iyi hissediyorum, şimdiye kadar öyle değil *anlama* anlamlı bir anlamda ne söylediği. Büyük veri kümeleri üzerinde eğitilmesinden itibaren, bu sözcükleri gerçek dünyadaki hiçbir şeyle ilişkilendirmeden sözcükleri içeren kalıpları ve ilişkileri keşfeder. Örneğin, her kelimeyi belirli başka kelimelere ne kadar benzer olduğunu belirten bin sayılık bir listeyle temsil edebilir. Bundan sonra, bundan “kral” ve “kraliçe” arasındaki farkın “koca” ve “karı” arasındaki farka benzer olduğu sonucuna varabilir - ama yine de erkek ya da kadın olmanın ne anlama geldiğine dair hiçbir fikri yoktur ya da uzay, zaman ve madde ile fiziksel gerçeklik gibi bir şey.

Turing testi temelde aldatma ile ilgili olduğundan, insan saflığını gerçek yapay zekadan daha fazla test ettiği için eleştirildi. Aksine, rakip bir test olarak adlandırılan *Winograd Şema Mücadelesi* Doğrudan juguler için gider, mevcut derin öğrenme sistemlerinin eksik olma eğiliminde olduğu sağduyu anlayışına odaklanır. Biz insanlar bir zamirin neyi ifade ettiğini anlamak için bir cümleyi çözümlerken rutin olarak gerçek dünya bilgisini kullanırız. Örneğin, tipik bir Winograd yarışması burada "onlar" ın neyi ifade ettiğini sorar:

1. "Belediye meclis üyeleri, şiddetten korktukları için göstericilere izin vermeyi reddetti."

2. "Belediye meclis üyeleri, şiddeti savundukları için göstericilere izin vermeyi reddetti."

Bu tür soruları yanıtlamak için yıllık bir AI yarışması var ve AI'lar hala perişan bir şekilde performans gösteriyor.[7] Bu kesin meydan okuma, neyin neyi ifade ettiğini anlamak, yukarıdaki örneğimde İspanyolca'yı Çince ile değiştirdiğimde GoogleTranslate'i bile baltaladı:

Ancak AI, 2016'da büyük bir aradan sonra neredeyse hiç dil kullanmadan beni yakaladı, AI sistemini Google Brain tarafından geliştirilenden daha çevirebildim.

takım.

Lütfen kendin dene https://translate.google.com Artık kitabı okuyorsunuz ve Google'ın yapay zekasının gelişip gelişmediğini görün! Bir dünya modeli içeren bir dil işleme yapay zekası oluşturmak için derin yinelenen sinir ağlarını GOFAI ile evlendirmek için umut verici yaklaşımlar olduğundan, sahip olma şansı çok yüksek.

Fırsatlar ve Zorluklar

Yapay zeka birçok önemli cephede hızla ilerlediğinden, bu üç örnek açıkça sadece bir örnekleyiciydi. Dahası, bu örneklerde sadece iki şirketten bahsetmiş olsam da, üniversitelerdeki ve diğer şirketlerdeki rakip araştırma grupları genellikle çok geride değildi. Apple, Baidu, DeepMind, Facebook, Google, Microsoft ve diğerleri öğrencileri, postdocları ve fakülteyi uzaklaştırmak için kazançlı teklifler kullandığından, dünyanın dört bir yanındaki bilgisayar bilimi bölümlerinde yüksek bir emme sesi duyulabilir.

Yapay zekanın tarihini ara sıra atılımla noktalanan durgunluk dönemleri olarak görmeye verdiğim örnekler tarafından yanıltılmamak önemlidir. Benim bakış açımdan, bunun yerine uzun süredir oldukça istikrarlı bir ilerleme görüyorum - medya, yeni bir hayal gücü kapma uygulaması veya yararlı bir ürün sağlama eşiğini her aştığında bir atılım olarak rapor ediyor. Bu nedenle, canlı AI ilerlemesinin uzun yıllar devam edeceğini düşünüyorum. Dahası, son bölümde gördüğümüz gibi, yapay zeka çoğu görevde insan yeteneklerini eşleştirene kadar bu ilerlemenin devam etmemesinin temel bir nedeni yok.

Hangisi soruyu gündeme getiriyor: Bu bizi nasıl etkileyecek? Yakın vadeli yapay zeka ilerlemesi, insan olmanın anlamını nasıl değiştirecek? Yapay zekanın hedeflerden, genişlikten, sezgiden, yaratıcılıktan veya dilden tamamen yoksun olduğunu iddia etmenin giderek zorlaştığını gördük - birçoğunun insan olmanın merkezinde olduğunu düşündüğü özellikler. Bu, yakın vadede bile, herhangi bir AGI bizi tüm görevlerde eşleştirmeden çok önce, yapay zekanın kendimizi nasıl gördüğümüz, yapay zeka ile tamamlandığında neler yapabileceğimiz ve ne zaman yaparak para kazanabileceğimiz üzerinde dramatik bir etkiye sahip olabileceği anlamına gelir. AI ile rekabet etmek. Bu etki daha iyi mi yoksa daha kötü mü olacak? Bu, kısa vadede ne gibi fırsatlar ve zorluklar sunacak?

Uygarlıkla ilgili sevdiğimiz her şey insan zekasının bir ürünüdür, bu yüzden onu yapay zeka ile büyütebilirsek, açıkça yaşamı daha da iyi hale getirme potansiyeline sahibiz. AI'daki mütevazı ilerleme bile bilim ve teknolojide büyük gelişmelere ve buna bağlı olarak kaza, hastalık, adaletsizlik, savaş, angarya ve yoksullukta azalma anlamına gelebilir. Ancak yapay zekanın bu faydalarından yeni sorunlar yaratmadan yararlanmak için birçok önemli soruyu yanıtlamamız gerekiyor. Örneğin:

1. İstediğimizi çökmeden, arızalanmadan veya saldırıya uğramadan yapmaları için gelecekteki yapay zeka sistemlerini bugünkünden daha sağlam hale nasıl getirebiliriz?
2. Hukuk sistemlerimizi daha adil ve verimli olacak ve hızla değişen dijital ortama ayak uyduracak şekilde nasıl güncelleyebiliriz?
3. Ölümcül otonom silahlarda kontrolden çıkmış bir silahlanma yarışını tetiklemeden silahları nasıl daha akıllı ve masum sivilleri öldürmeye daha az eğilimli hale getirebiliriz?
4. İnsanları gelirden veya amaçtan yoksun bırakmadan otomasyon yoluyla refahımızı nasıl artırabiliriz?

Bu bölümün geri kalanını bu soruların her birini sırayla incelemeye ayıralım. Bu dört kısa vadeli soru, sırasıyla bilgisayar bilimcileri, hukuk akademisyenleri, askeri stratejistler ve iktisatçılara yöneliktir. Ancak, ihtiyaç duyduğumuz cevaplara ihtiyacımız olduğu anda ulaşmamıza yardımcı olmak için herkesin bu sohbete katılması gerekiyor, çünkü göreceğimiz gibi, zorluklar hem uzmanlıklar hem de uluslar arası tüm geleneksel sınırları aşıyor.

Hatalar ve Sağlam AI

Bilgi teknolojisinin bilimden finansa, üretim, ulaşım, sağlık hizmetleri, enerji ve iletişime kadar insan girişimimizin hemen hemen her sektörü üzerinde büyük bir olumlu etkisi olmuştur ve bu etki, yapay zekanın getirme potansiyeline sahip olduğu ilerlemeye kıyasla çok düşüktür. Ancak teknolojiye ne kadar güvenirsek, sağlam ve güvenilir olması, yapmasını istediğimiz şeyi yapması o kadar önemli hale geliyor.

İnsanlık tarihi boyunca, teknolojimizi faydalı kılmak için aynı denenmiş ve doğru yaklaşıma güvendik: hatalardan ders almak. Yangını icat ettik, defalarca karıştırdık ve sonra yangın söndürücüyü, yangın çıkışını, yangın alarmını ve itfaiyeyi icat ettik. Otomobili icat ettik, defalarca kaza yaptık ve ardından emniyet kemerleri, hava yastıkları ve kendi kendine giden arabaları icat ettik. Şimdiye kadar, teknolojilerimiz tipik olarak, yararları nedeniyle zararlarının ağır basmasına yetecek kadar az sayıda ve sınırlı kazaya neden oldu. Her zamankinden daha güçlü bir teknoloji geliştirdikçe, kaçınılmaz olarak tek bir kazanın bile tüm faydalardan daha ağır basacak kadar yıkıcı olabileceği bir noktaya ulaşacağız. Bazıları tesadüfi küresel nükleer savaşın böyle bir örnek teşkil edeceğini iddia ediyor. Diğerleri biyomühendislik ürünü bir pandeminin uygun olabileceğini savunuyor ve bir sonraki bölümde, Gelecekteki yapay zekanın insan neslinin tükenmesine neden olup olamayacağı konusundaki tartışmaları inceleyeceğiz. Ancak önemli bir sonuca ulaşmak için bu tür uç örnekleri düşünmemize gerek yok: teknoloji güçlendikçe, güvenlik mühendisliği için deneme-yanılma yaklaşımına daha az güvenmeliyiz. Diğer bir deyişle, *reaktif olmaktan çok proaktif olmalıyız,* Kazaların bir kez bile olmasını önlemeyi amaçlayan güvenlik araştırmalarına yatırım yapmak. Toplumun nükleer reaktör güvenliğine fare kapanı güvenliğinden daha fazla yatırım yapmasının nedeni budur.

Bu aynı zamanda, 1. bölümde gördüğümüz gibi, Porto Riko konferansında yapay zeka güvenliği araştırmalarına halkın büyük ilgisinin olmasının sebebidir. Bilgisayarlar ve AI sistemleri her zaman çöktü, ancak bu sefer farklı: AI yavaş yavaş gerçek dünyaya giriyor ve elektrik şebekesini, borsayı veya bir nükleer silah sistemini çökertmesi sadece bir sıkıntı değil. Bu bölümün geri kalanında, sizi şu anki AI güvenliği tartışmasına hâkim olan ve halen devam eden dört ana teknik AI güvenlik araştırması alanını tanıtmak istiyorum.

dünya çapında takip edildi: *doğrulama, doğrulama, güvenlik* ve *kontrol.* *[1] İçin

işlerin fazla zorlaşmasını ve kurumasını önleyin, bunu bilgi teknolojisinin farklı alanlardaki geçmiş başarılarını ve başarısızlıklarını, ayrıca onlardan öğrenebileceğimiz değerli dersleri ve bunların ortaya çıkardığı zorlukları araştırarak yapalım.

Bu öykülerin çoğu eski olsa da, neredeyse hiç kimsenin yapay zeka olarak adlandırmayacağı ve çok az can kaybına neden olan düşük teknolojili bilgisayar sistemlerini içeriyor olsa da, yine de bize güvenli ve güçlü bir geleceğin yapay zekasını tasarlamak için değerli dersler öğrettiklerini göreceğiz. arızaları gerçekten felaket olabilecek sistemler.

Uzay Keşfi için AI

Kalbime yakın bir şeyle başlayalım: uzay araştırması. Bilgisayar teknolojisi, insanları Ay'a uçurmamızı ve Güneş Sistemimizin tüm gezegenlerini keşfetmek için insansız uzay araçları göndermemizi, hatta Satürn'ün ayı Titan'a ve bir kuyruklu yıldıza inmemizi sağladı. Bölüm 6'da keşfedeceğimiz gibi, gelecekteki yapay zeka diğer güneş sistemlerini ve galaksileri keşfetmemize yardımcı olabilir - eğer hatasızsa. Haziranda

4 Ocak 1996'da, Dünya'nın manyetosferini araştırmayı uman bilim adamları, Avrupa Uzay Ajansı'ndan bir Ariane 5 roketi yaptıkları bilimsel aletlerle gökyüzüne kükrerken sevinçle alkışladılar. Otuz yedi saniye sonra, roket yüzlerce yıla mal olan bir havai fişek gösterisinde patlarken gülümsemeleri kayboldu.

milyonlarca dolar. [8] Nedenin, kendisine ayrılan 16 bite sığamayacak kadar büyük bir sayıyı manipüle eden hatalı yazılım olduğu bulundu. [9] İki yıl sonra, NASA'nın Mars Climate Orbiter'ı yanlışlıkla Kızıl Gezegen'in atmosferine girdi ve parçalandı çünkü yazılımın iki farklı parçası güç için farklı birimler kullandı ve roket motoru itme kontrolünde% 445 hataya neden oldu. [10] Bu, NASA'nın ikinci süper pahalı hatasıydı: Venüs'e yaptıkları Mariner 1 görevi, uçuş kontrol yazılımının yanlış bir noktalama işaretiyle engellendiği 22 Temmuz 1962'de Cape Canaveral'dan fırlatıldıktan sonra patladı. [11] Sanki sadece batılıların uzaya böcek fırlatma sanatında ustalaşmadığını göstermek için, Sovyet Phobos 1 görevi 2 Eylül 1988'de başarısız oldu. Bu, Mars'a bir iniş uçağı yerleştirme gibi muhteşem bir hedefle fırlatılan en ağır gezegenler arası uzay aracıydı. moon Phobos - eksik bir kısa çizgi, Mars'a giderken uzay aracına "görev sonu" komutunun gönderilmesine ve tüm sistemlerini kapatmasına neden olunca hepsi engellendi. [12]

Bu örneklerden öğrendiklerimiz, bilgisayar bilimcilerin dediği şeyin önemi *doğrulama:* Yazılımın beklenen tüm gereksinimleri tam olarak karşılamasını sağlamak. Ne kadar çok hayat ve kaynak söz konusu olursa, yazılımın amaçlandığı gibi çalışacağına dair o kadar yüksek güven isteriz. Neyse ki, AI, doğrulama sürecini otomatikleştirmeye ve iyileştirmeye yardımcı olabilir. Örneğin, eksiksiz, genel amaçlı bir işletim sistemi çekirdeği adı verilir *seL4* son zamanlarda

Çökmelere ve güvenli olmayan işlemlere karşı güçlü bir garanti vermek için matematiksel olarak resmi bir spesifikasyona göre kontrol edildi: Microsoft Windows ve Mac OS'nin çan ve ıslıklarıyla henüz gelmese de, size sevgiyle bilinen şeyleri vermeyeceğinden emin olabilirsiniz. "mavi ölüm ekranı" olarak veya

"Kıyametin çıkrık çarkı." ABD Savunma Gelişmiş Araştırma Projeleri Ajansı (DARPA), kanıtlanabilir şekilde güvenli olan HACMS (yüksek güvenceli siber askeri sistemler) adı verilen bir dizi açık kaynaklı yüksek güvence aracının geliştirilmesine sponsor olmuştur. Önemli bir zorluk, bu tür araçları yeterince güçlü ve kullanımı kolay hale getirerek geniş çapta konuşlandırılmalarını sağlamaktır. Diğer bir zorluk ise, yazılım robotlara ve yeni ortamlara girdikçe ve geleneksel önceden programlanmış yazılımların yerini öğrenmeye devam eden ve böylece davranışlarını değiştiren yapay zeka sistemlerine bıraktıkça, doğrulama görevinin kendisi daha da zorlaşacak olmasıdır.

Finans için AI

Finans, bilgi teknolojisi tarafından dönüştürülmüş, kaynakların ışık hızında verimli bir şekilde yeniden tahsis edilmesine ve ipoteklerden başlangıç şirketlerine kadar her şey için uygun maliyetli finansman sağlayan başka bir alandır. Yapay zeka alanındaki ilerlemenin, finansal ticaretten gelecekte büyük kâr fırsatları sunması muhtemeldir: çoğu borsa alım / satım kararları artık otomatik olarak bilgisayarlar tarafından alınmaktadır ve mezun olan MIT öğrencilerim rutin olarak, algoritmik ticareti geliştirmek için astronomik başlangıç maaşlarından etkilenirler.

Doğrulama, finansal yazılımlar için de önemli, Amerikan firması Knight Capital'in 1 Ağustos 2012'de 440 dolar kaybederek zor yoldan öğrendiği doğrulanmamış ticaret yazılımı dağıtıldıktan sonra kırk beş dakika içinde milyon. [13] 6 Mayıs 2010'da yaşanan trilyon dolarlık "Flaş Çöküşü" farklı bir nedenden ötürü dikkate değerdi. Piyasalar istikrara kavuşmadan önce yaklaşık yarım saat boyunca büyük aksamalara neden olmasına rağmen, Procter &
Bir kuruş ile 100.000 dolar arasında değişen kumar, [14] sorun, doğrulamanın önleyebileceği hatalar veya bilgisayar arızalarından kaynaklanmıyordu. Bunun yerine, beklentilerin ihlal edilmesinden kaynaklanıyordu: Birçok şirketin otomatik alım satım programları, kendilerini varsayımlarının geçerli olmadığı beklenmedik bir durumda çalışırken buldular - örneğin, bir borsa bilgisayarının bir hisse senedinin fiyatı olduğunu rapor ederse varsayımı bir sent, sonra o hisse senedi gerçekten bir sent değerindeydi.

Flaş çökmesi, bilgisayar bilimcilerinin dediği şeyin önemini gösteriyor *doğrulama:* doğrulama ise "Sistemi doğru kurdum mu?" sorusunu sorar, doğrulama "Doğru sistemi kurdum mu?" diye soruyor * [2] Örneğin, sistem her zaman geçerli olmayabilecek varsayımlara mı güveniyor? Öyleyse, belirsizlikle daha iyi başa çıkmak için nasıl geliştirilebilir?

Üretim için AI

Söylemeye gerek yok, AI, hem verimliliği hem de hassasiyeti artıran robotları kontrol ederek üretimi iyileştirmek için büyük bir potansiyele sahip. Sürekli gelişen 3-D yazıcılar artık ofis binalarından yazıcıya kadar her şeyin prototiplerini yapabilir.

tuz tanesinden daha küçük mikromekanik cihazlar. [15] Devasa endüstriyel robotlar arabalar ve uçaklar üretirken, uygun fiyatlı bilgisayar kontrollü değirmenler, torna tezgahları, kesiciler ve benzerleri yalnızca fabrikalara değil aynı zamanda yerel meraklıların fikirlerini bir anda somutlaştırdıkları taban "yapımcı hareketine" de güç veriyor.

dünya çapında binlerce topluluk tarafından işletilen "fabrika laboratuvarı". [16] Ancak çevremizde ne kadar çok robot varsa, yazılımlarını doğrulamamız ve doğrulamamız o kadar önemli hale geliyor. Bir robot tarafından öldürüldüğü bilinen ilk kişi, Michigan, Flat Rock'taki bir Ford fabrikasında işçi olan Robert Williams'tı. 1979'da, bir depolama alanından parçaları alması gereken bir robot arızalandı ve parçaları kendisi almak için alana tırmandı. Robot sessizce çalışmaya başladı ve kafasını kırdı, iş arkadaşlarına gelene kadar otuz dakika devam etti.

ne olduğunu keşfetti. [17] Bir sonraki robot kurbanı, Japonya, Akashi'deki Kawasaki fabrikasında bakım mühendisi olan Kenji Urada idi. 1981'de kırık bir robot üzerinde çalışırken, yanlışlıkla açma düğmesine bastı ve ezilerek öldü.

robotun hidrolik kolu ile. [18] 2015 yılında, Volkswagen'in Almanya, Baunatal'daki üretim tesislerinden birinde yirmi iki yaşındaki bir müteahhit, otomobil parçalarını kapmak ve işlemek için bir robot kurmaya çalışıyordu. Bir şeyler ters gitti, robotun onu yakalayıp bir metale karşı ezmesine neden oldu

tabak. [19]

Bu kazalar trajik olsa da, tüm endüstriyel kazaların çok küçük bir kısmını oluşturduklarını unutmamak önemlidir. Üstelik endüstriyel kazalar *azaldı* teknoloji geliştikçe artmaktansa,

Amerika Birleşik Devletleri'nde 1970'te yaklaşık 14.000 ölüm, 2014'te 4.821'e. [20] Yukarıda bahsedilen üç kaza, aksi takdirde aptal olan makinelere istihbarat eklemenin, robotların insanların etrafında daha dikkatli olmayı öğrenmesini sağlayarak endüstriyel güvenliği daha da artırabileceğini gösteriyor. Üç kazanın tümü daha iyi bir doğrulama ile önlenebilirdi: Robotlar, hatalar veya kötü niyet nedeniyle değil, kişinin orada olmadığı veya kişinin bir otomobil parçası olduğu gibi geçersiz varsayımlar yaptıkları için zarara neden oldu.



Şekil 3.3: Geleneksel endüstriyel robotlar pahalı ve programlanması zor olsa da, programlama deneyimi olmayan işçilerden ne yapılacağını öğrenebilen daha ucuz AI destekli robotlara doğru bir eğilim var.

Ulaşım için AI

Yapay zeka, üretimde birçok hayat kurtarabilirse de, ulaşımda potansiyel olarak daha da fazla tasarruf sağlayabilir. Sadece araba kazaları 2015 yılında 1,2 milyon can aldı ve uçak, tren ve tekne kazaları birlikte binlerce kişiyi daha öldürdü. Yüksek güvenlik standartlarına sahip Amerika Birleşik Devletleri'nde geçen yıl motorlu taşıt kazaları yaklaşık 35.000 kişinin ölümüne neden oldu - bu, tüm endüstriyel kazalardan yedi kat daha fazla.

kombine. [21] Bu konuyla ilgili olarak Austin, Texas'ta Yapay Zekayı Geliştirme Derneği'nin 2016 yıllık toplantısında bir panel tartışması yaptığımızda, İsrailli bilgisayar bilimcisi Moshe Vardi bu konuda oldukça duygusallaştı ve sadece *abilir* AI, yol ölümlerini azaltır, ancak *zorunlu:* "Bu ahlaki bir zorunluluk!" diye haykırdı. Neredeyse tüm araba kazalarının nedeni insan hatasından kaynaklandığı için, yapay zeka ile çalışan kendi kendine giden arabaların yol ölümlerinin en az% 90'ını ortadan kaldırabileceğine inanılıyor ve bu iyimserlik, kendi kendine giden arabaları yollara çıkarmaya yönelik büyük ilerlemeyi körüklüyor. . Elon Musk, gelecekteki sürücüsüz arabaların sadece daha güvenli olmayacağını, aynı zamanda Uber ve Lyft ile rekabet ederek ihtiyaç duyulmadıklarında sahipleri için para kazanacağını öngörüyor.

Şu ana kadar, sürücüsüz arabalar gerçekten de insan sürücülerden daha iyi bir güvenlik siciline sahip ve meydana gelen kazalar, onaylamanın önemi ve zorluğunun altını çiziyor. Google'ın sürücüsüz arabasının neden olduğu ilk çamurluk bükücü 14 Şubat 2016'da gerçekleşti, çünkü bir otobüs hakkında yanlış bir varsayımda bulundu: otomobil önünden çekildiğinde sürücünün yol vereceği. 7 Mayıs 2016'da otoyoldan geçen bir kamyonun römorkuna çarpan sürücüsüz Tesla'nın yol açtığı ilk ölümcül kazaya iki kötü neden oldu.

varsayımlar: [22] römorkun parlak beyaz tarafının sadece parlak gökyüzünün bir parçası olduğunu ve sürücünün (iddia edilen bir Harry Potter
film) dikkat ediyordu ve bir şeyler ters giderse müdahale edecekti. * [3]

Ancak bazen iyi bir doğrulama ve onaylama, kazaları önlemek için yeterli değildir, çünkü aynı zamanda iyi *kontrol:* bir insan operatörünün sistemi izleme ve gerekirse davranışını değiştirme yeteneği. Bunun için *döngüdeki insan* sistemlerin iyi çalışması için insan-makine iletişiminin etkili olması çok önemlidir. Bu ruhla, arabanızın bagajını yanlışlıkla açık bırakırsanız, gösterge panelinizdeki kırmızı ışık sizi rahatlıkla uyaracaktır. Buna karşılık, İngiliz araba feribotu *Özgür Teşebbüsün Habercisi* 6 Mart 1987'de Zeebrugge limanından, pruva kapıları açık olarak ayrıldı, hiçbir uyarı ışığı veya başka bir şey yoktu

Kaptan için gözle görülür uyarı ve limandan ayrıldıktan kısa bir süre sonra feribot alabora oldu ve 193 kişi öldü. [23]

Daha iyi makine-insan iletişimiyle önlenebilecek başka bir trajik kontrol hatası, Air France Flight 447'nin Atlantik Okyanusu'na düştüğü 1 Haziran 2009 gecesi meydana geldi ve gemideki 228 kişi öldü. Resmi kaza raporuna göre, "mürettebat, durduklarını asla anlamadı ve sonuç olarak, çok geç olana kadar uçağın burnunu aşağı itmeyi içeren bir kurtarma manevrası uygulamadı". Uçuş güvenliği uzmanları, kokpitte pilotlara burnun çok yukarı dönük olduğunu gösteren bir “hücum açısı” göstergesi olsaydı kazanın önlenebileceğini tahmin ettiler. [24]

Air Inter Flight 148, 20 Ocak 1992'de Fransa'nın Strazburg yakınlarındaki Vosges Dağları'na çarparak 87 kişiyi öldürdüğünde, bunun nedeni makine-insan iletişimi eksikliği değil, kafa karıştırıcı bir kullanıcı arayüzü idi. Pilotlar, 3.3 derecelik bir açıyla inmek istedikleri için bir tuş takımıyla "33" e girdiler, ancak otomatik pilot bunu, farklı bir modda olduğu için dakikada 3.300 fit olarak yorumladı ve ekran modu gösteremeyecek kadar küçüktü ve pilotların hatalarını fark etmelerine izin verin.

Enerji için AI

Bilgi teknolojisi, dünyanın elektrik şebekelerinde üretim ve tüketimi dengeleyen sofistike algoritmalar ve enerji santrallerinin güvenli ve verimli çalışmasını sağlayan sofistike kontrol sistemleriyle güç üretimi ve dağıtımı için harikalar yarattı. Gelecekteki yapay zeka ilerlemesi, büyük olasılıkla "akıllı şebekeyi" daha da akıllı hale getirecek ve değişen arz ve talebe en uygun şekilde, çatı üstü güneş panelleri ve ev pil sistemleri seviyesine kadar bile en iyi şekilde adapte olacak. Ancak 14 Ağustos 2003 Perşembe günü, Amerika Birleşik Devletleri ve Kanada'da çoğu günlerce güçsüz kalan yaklaşık 55 milyon insan için ışıklar söndü. Burada da birincil nedenin başarısız makine-insan iletişimi olduğu belirlendi:

kontrol. [25]

28 Mart 1979'da Pennsylvania'daki Three Mile Adası'ndaki bir reaktördeki kısmi nükleer erime, yaklaşık bir milyar dolarlık temizlik maliyetine ve nükleer enerjiye karşı büyük bir tepkiye yol açtı. Nihai kaza raporu, zayıf bir kullanıcının neden olduğu kafa karışıklığı da dahil olmak üzere birçok katkıda bulunan faktörü belirledi.

arayüz. [26] Özellikle, operatörlerin düşündüğü uyarı ışığı, güvenlik açısından kritik bir vananın açık mı yoksa kapalı mı olduğunu gösteriyordu, yalnızca vanayı kapatmak için bir sinyal gönderilip gönderilmediğini gösteriyordu - bu nedenle operatörler vananın açık kaldığının farkında değildi.

Bu enerji ve ulaşım kazaları bize, yapay zekayı her zamankinden daha fazla fiziksel sistemden sorumlu tuttuğumuzda, yalnızca makinelerin kendi başlarına iyi çalışmasını sağlamak için değil, aynı zamanda makinelerin insan denetleyicileriyle etkili bir şekilde işbirliği yapmasını sağlamak için ciddi araştırma çabaları göstermemiz gerektiğini öğretir. . Yapay zeka daha akıllı hale geldikçe, bu sadece bilgi paylaşımı için iyi kullanıcı arayüzleri oluşturmayı değil, aynı zamanda insan-bilgisayar ekipleri içinde görevlerin en iyi şekilde nasıl tahsis edileceğini (örneğin, kontrolün aktarılması gereken durumların belirlenmesi ve insan yargısını verimli bir şekilde uygulamak) içerecektir. önemsiz bir bilgi seliyle insan kontrolörlerinin dikkatini dağıtmak yerine en yüksek değerli kararlara.

Sağlık Hizmetleri için AI

AI, sağlık hizmetlerini iyileştirmek için büyük bir potansiyele sahiptir. Tıbbi kayıtların dijitalleştirilmesi, halihazırda doktorların ve hastaların daha hızlı ve daha iyi kararlar almalarına ve dijital görüntülerin teşhisinde dünya çapındaki uzmanlardan anında yardım almalarına olanak sağlamıştır. Aslında, bilgisayarla görme ve derin öğrenmedeki hızlı ilerleme göz önüne alındığında, bu tür teşhisi yapmak için en iyi uzmanlar yakında AI sistemleri olabilir. Örneğin, 2015 Hollanda çalışması, prostat kanserinin bilgisayarla teşhis edildiğini gösterdi.

manyetik rezonans görüntüleme (MRI) insan radyologlarınınki kadar iyiydi, [27]
ve 2016 Stanford çalışması, yapay zekanın akciğer kanserini
mikroskop görüntüleri insan patologlarından bile daha iyi. [28] Makine öğrenimi genler, hastalıklar ve tedavi yanıtları arasındaki ilişkileri ortaya çıkarmaya yardımcı olabilirse, kişiselleştirilmiş tıpta devrim yaratabilir, çiftlik hayvanlarını daha sağlıklı hale getirebilir ve daha dayanıklı mahsuller sağlayabilir. Dahası, robotlar, gelişmiş yapay zeka kullanmadan bile insanlardan daha doğru ve güvenilir cerrahlar olma potansiyeline sahip. Son yıllarda çok çeşitli robotik ameliyatlar başarıyla gerçekleştirildi ve sıklıkla hassaslık, minyatürleştirme ve daha küçük kesilerle kan kaybının azalmasına, daha az ağrıya ve daha kısa iyileşme süresine yol açtı.

Ne yazık ki, sağlık sektöründe de sağlam yazılımın önemi hakkında acı veren dersler alındı. Örneğin, Kanada yapımı Therac-25 radyasyon terapi makinesi kanser hastalarını iki farklı modda tedavi etmek için tasarlandı: ya düşük güçlü bir elektron ışınıyla ya da hedefte tutulan yüksek güçlü bir megavolt X ışınları ışınıyla özel bir kalkanla. Ne yazık ki, doğrulanmamış buggy yazılımı bazen teknisyenlerin düşük güçlü ışını yönettiklerini düşündüklerinde megavolt ışını göndermelerine neden oldu.

ve kalkan olmadan, birkaç hastanın hayatına mal oldu. [29]
Panama'daki Ulusal Onkoloji Enstitüsü'ndeki aşırı radyasyon dozlarından çok daha fazla hasta öldü, burada radyoaktif kobalt-60 kullanan radyoterapi ekipmanı, 2000 ve 2001 yıllarında aşırı maruz kalma sürelerine programlandı.
düzgün doğrulanmamış kafa karıştırıcı kullanıcı arayüzü. [30] Yakın tarihli bir rapora göre, [31] robotik cerrahi kazaları, Amerika Birleşik Devletleri'nde 2000 ve 2013 yılları arasında 144 ölüm ve 1.391 yaralanmayla ilişkilendirildi; yalnızca elektrik arkları ve hastaya düşen yanmış veya kırılmış alet parçaları gibi donanım sorunları değil, aynı zamanda yazılım sorunları gibi yaygın sorunlar da var. kontrolsüz hareketler ve kendiliğinden kapanma.

İyi haber şu ki, raporun kapsadığı neredeyse iki milyon robotik ameliyatın geri kalanı sorunsuz geçti ve robotlar ameliyatı daha az güvenli değil, daha çok yapıyor gibi görünüyor. ABD hükümeti tarafından yapılan bir araştırmaya göre, kötü hastane bakımı sadece Amerika Birleşik Devletleri'nde yılda 100.000'den fazla ölüme katkıda bulunur, [32] bu nedenle tıp için daha iyi yapay zeka geliştirmenin ahlaki zorunluluğu, tartışmasız kendi kendine giden arabalardan daha güçlüdür.

İletişim için AI

İletişim endüstrisi, muhtemelen bilgisayarların şimdiye kadar en büyük etkiye sahip olduğu sektördür. 1950'lerde bilgisayarlı telefon santrallerinin, altmışlarda internetin ve

1989, milyarlarca insan artık iletişim kurmak, alışveriş yapmak, haber okumak, film izlemek veya oyun oynamak için çevrimiçi oluyor, dünyanın bilgisine yalnızca bir tık uzaklıkta ve genellikle ücretsiz olarak alışkın. Ortaya çıkan *nesnelerin interneti* çiftlik hayvanlarında lambalardan, termostatlardan ve dondurucilardan biyoçip transponderlere kadar her şeyi çevrimiçi hale getirerek daha fazla verimlilik, doğruluk, rahatlık ve ekonomik fayda vaat ediyor.

Dünyayı birbirine bağlamadaki bu muhteşem başarılar, bilgisayar bilimcilerini dördüncü bir zorluğu da beraberinde getirdi: yalnızca doğrulama, onaylama ve kontrolü değil, aynı zamanda *güvenlik* kötü amaçlı yazılımlara ("kötü amaçlı yazılım") ve saldırılara karşı. Yukarıda belirtilen sorunların tamamı kasıtsız hatalardan kaynaklanırken, güvenlik *kasıtlı suistimal.* Medyanın dikkatini çeken ilk kötü amaçlı yazılım, 2 Kasım 1988'de ortaya çıkan ve UNIX işletim sistemindeki hataları kullanan Morris solucanıydı. İddiaya göre, kaç bilgisayarın çevrimiçi olduğunu saymak için yanlış bir girişimdi ve o zamanlar interneti oluşturan 60.000 bilgisayarın yaklaşık% 10'una bulaşıp çökmesine rağmen, bu, yaratıcısı Robert Morris'in sonunda bir MIT'de bilgisayar bilimi alanında kadrolu profesörlük.

Diğer kötü amaçlı yazılımlar, yazılımlardaki değil, insanlardaki güvenlik açıklarından yararlanır. Mayısta 5, 2000, sanki benim doğum günümü kutlamak istercesine, insanlar tanıdıklarından ve meslektaşlarından ve istemeden "LOVE-LETTER-FOR-YOU.txt.vbs" ekine tıklayan Microsoft Windows kullanıcılarından "ILOVEYOU" konu satırını içeren e-postalar aldı. bilgisayarlarına zarar veren ve e-postayı adres defterlerindeki herkese gönderen bir komut dosyası başlattı. Filipinler'deki iki genç programcı tarafından yaratılan bu solucan, tıpkı Morris solucanının yaptığı gibi internetin yaklaşık% 10'unu enfekte etti, ancak o zamana kadar internet çok daha büyük olduğu için, tüm zamanların en büyük enfeksiyonlarından biri haline geldi. 50 milyondan fazla bilgisayar ve 5 milyar doların üzerinde hasara neden oluyor. Muhtemelen acı verici bir şekilde farkında olduğunuz gibi, internet, güvenlik uzmanlarının solucanlar, Truva atları olarak sınıflandırdığı sayısız bulaşıcı kötü amaçlı yazılım türünün istilasına uğramaya devam ediyor.

dosyalarınızı silmek, kişisel bilgilerinizi çalmak, sizi gözetlemek ve spam göndermek için bilgisayarınızı ele geçirmek için zararsız şaka mesajları gösterme.

Kötü amaçlı yazılım yapabildiği bilgisayarı hedef alırken, *hackerlar* belirli ilgilenilen hedeflere saldırmak — Target, TJ Maxx, Sony Pictures, Ashley Madison, Suudi petrol şirketi Aramco ve ABD Demokratik Ulusal Komitesi dahil olmak üzere son zamanlardaki yüksek profilli örnekler. Dahası, ganimetlerin daha da muhteşem olduğu görülüyor. Bilgisayar korsanları, 2008'de Heartland Payment Systems'dan 130 milyon kredi kartı numarası ve diğer hesap bilgilerini çaldı ve bu bilgileri ihlal etti

bir milyardan fazla (!) Yahoo! 2013 yılında e-posta hesapları. [33] ABD Hükümeti Personel Yönetimi Ofisi'nin 2014'te yaptığı bir hack, iddiaya göre en yüksek güvenlik izinlerine sahip çalışanlar ve gizli ajanların parmak izlerini içeren 21 milyondan fazla kişinin personel kayıtlarını ve iş başvurusu bilgilerini ihlal etti.

Sonuç olarak,% 100 güvenli ve hacklenemez olduğu iddia edilen yeni bir sistem hakkında her okuduğumda gözlerimi deviriyorum. Yine de, “hacklenemez”, gelecekteki AI sistemlerini, diyelim ki kritik altyapı veya silah sistemlerinden sorumlu hale getirmeden önce olması gereken şeydir, bu nedenle AI'nın toplumdaki artan rolü, bilgisayar güvenliği için riskleri artırmaya devam ediyor. Bazı bilgisayar korsanları, yeni çıkan yazılımdaki insan saflığından veya karmaşık güvenlik açıklarından yararlanırken, diğerleri, utanç verici derecede uzun bir süre boyunca fark edilmeden kalan basit hatalardan yararlanarak uzak bilgisayarlara yetkisiz giriş yapılmasını sağlar. Bilgisayarlar arasında güvenli iletişim için en popüler yazılım kitaplıklarından birinde "Heartbleed" hatası 2012'den 2014'e kadar sürdü ve "Bashdoor" hatası, 1989'dan 2014'e kadar Unix bilgisayarların işletim sistemine dahil edildi.

Ne yazık ki, daha iyi AI sistemleri, yeni güvenlik açıkları bulmak ve daha karmaşık hackler gerçekleştirmek için de kullanılabilir. Örneğin, bir gün sizi kişisel bilgilerinizi ifşa etmeye ikna etmeye çalışan alışılmadık şekilde kişiselleştirilmiş bir "kimlik avı" e-postası aldığınızı hayal edin. Arkadaşınızın hesabından, onu hackleyen ve onun kimliğine bürünen, diğer gönderilen e-postalarının analizine dayanarak yazma stilini taklit eden ve diğer kaynaklardan sizin hakkınızda birçok kişisel bilgi içeren bir AI tarafından gönderilir. Buna kanabilir misin? Ya kimlik avı e-postası kredi kartı şirketinizden geliyor gibi görünüyorsa ve ardından yapay zeka tarafından oluşturulmuş olduğunu söyleyemeyeceğiniz dost canlısı bir insan sesinden gelen bir telefon araması geliyorsa? Hücum ve savunma arasında devam eden bilgisayar güvenliği silahlanma yarışında, savunmanın kazandığına dair şimdiye kadar çok az gösterge var.

Kanunlar

Biz insanlar, diğer tüm türleri bastıran ve işbirliği yapma yeteneğimiz sayesinde Dünya'yı fetheden sosyal hayvanlarız. İşbirliğini teşvik etmek ve kolaylaştırmak için yasalar geliştirdik, böylece AI yasal ve yönetişim sistemlerimizi geliştirebilirse, o zaman her zamankinden daha başarılı bir şekilde işbirliği yapmamızı sağlayarak içimizdeki en iyiyi ortaya çıkarabilir. Ve burada hem yasalarımızın nasıl uygulandığı hem de nasıl yazıldığı konusunda pek çok gelişme fırsatı var, o halde ikisini de sırayla inceleyelim.

Ülkenizdeki mahkeme sistemi dendiğinde aklınıza gelen ilk dernekler hangileridir? Uzun gecikmeler, yüksek maliyetler ve ara sıra yaşanan adaletsizlikler ise, o zaman yalnız değilsiniz. İlk düşünceleriniz yerine “verimlilik” ve “adalet” olsaydı harika olmaz mıydı? Yasal süreç soyut olarak bir hesaplama, kanıtlar ve yasalar hakkında bilgi girme ve bir karar verme olarak görülebildiğinden, bazı bilim adamları bunu tamamen otomatik hale getirmeyi hayal ediyor.

robojudges: Önyargı, yorgunluk veya en son bilgi eksikliği gibi insan hatalarına boyun eğmeden aynı yüksek yasal standartları her yargıda yorulmadan uygulayan AI sistemleri.

Robojudges

Byron De La Beckwith Jr., 1994 yılında, 1963'te medeni haklar lideri Medgar Evers'e suikast yapmaktan suçlu bulundu, ancak iki ayrı tamamen beyaz Mississippi jürisi, fiziksel suçlara rağmen, cinayetten sonraki yıl onu mahkum edemedi.

kanıt esasen aynıydı. [34] Ne yazık ki, hukuk tarihi ten rengi, cinsiyet, cinsel yönelim, din, milliyet ve diğer faktörlere göre önyargılı yargılarla doludur. Robojudges, prensipte, tarihte ilk kez, herkesin yasa altında gerçekten eşit olmasını sağlayabilir: hepsi aynı olacak ve herkese eşit davranacak şekilde, yasayı gerçekten tarafsız bir şekilde şeffaf bir şekilde uygulayacak şekilde programlanabilirler.

Robojudges, kasıtlı olmaktan ziyade tesadüfi olan insan önyargılarını da ortadan kaldırabilir. Örneğin, 2012'de İsrailli yargıçların tartışmalı bir araştırması, aç olduklarında önemli ölçüde daha sert hükümler verdiklerini iddia etti: kahvaltıdan hemen sonra şartlı tahliye davalarının yaklaşık% 35'ini reddettiler, reddettiler.

öğle yemeğinden hemen önce% 85'in üzerinde. [35] İnsan yargıçların bir başka kusuru da, bir davanın tüm ayrıntılarını keşfetmek için yeterli zamana sahip olmayabilmeleridir. Aksine, robojudges yazılımdan biraz daha fazlasını içerdiğinden kolayca kopyalanabilir ve tüm bekleyen vakaların seri yerine paralel olarak işlenmesine izin verir, her vaka süresi boyunca kendi robojudge'ını alır. Son olarak, insan yargıçların çetrefilli patent ihtilaflarından en son adli bilime bağlı cinayet gizemlerine kadar her olası dava için gerekli tüm teknik bilgilere hakim olması imkansız olsa da, gelecekteki robojudların temelde sınırsız bellek ve öğrenme kapasitesi olabilir.

Bir gün, bu tür robojudges tarafsız, yetkin ve şeffaf olmaları nedeniyle hem daha verimli hem de daha adil olabilir. Verimlilikleri onları daha da adil hale getiriyor: hukuki süreci hızlandırarak ve bilgili avukatların sonucu çarpıtmasını zorlaştırarak, mahkemeler aracılığıyla adaleti sağlamayı önemli ölçüde daha ucuz hale getirebilirler. Bu, nakit sıkıntısı çeken bir bireyin veya bir avukat ordusuyla bir milyarder veya çok uluslu şirkete karşı galip gelen bir başlangıç şirketinin şansını büyük ölçüde artırabilir.

Öte yandan, robojudge'ların hataları varsa veya saldırıya uğrarsa ne olur? Her ikisi de otomatik oylama makinelerine zaten zarar verdi ve yıllarca parmaklıklar ardında veya bankada milyonlarca kişi söz konusu olduğunda, siber saldırıların teşvikleri daha da artıyor.

Yapay zeka, bir robojudge'ın yasal algoritmayı kullandığına güvenmemiz için yeterince sağlam hale getirilse bile, herkes onun muhakemesine saygı duyacak kadar mantıksal mantığını anladıklarını hissedecek mi? Bu zorluk, anlaşılmazlık pahasına geleneksel kolay anlaşılır yapay zeka algoritmalarından daha iyi performans gösteren sinir ağlarının son zamanlardaki başarısıyla daha da kötüleşiyor. Sanıklar bilmek isterse *neden* hüküm giymişlerdi, "sistemi birçok veriye göre eğittik ve buna karar verdik" ten daha iyi bir yanıt alma hakkına sahip olmaları gerekmez mi? Dahası, son araştırmalar, büyük miktarda mahkum verisi ile derin bir sinirsel öğrenme sistemi eğitirseniz, kimin suça dönme olasılığının yüksek olduğunu (ve bu nedenle şartlı tahliye reddedilmesi gerektiğini) insan hakimlerden daha iyi tahmin edebileceğini göstermiştir. Peki ya bu sistem, yeniden suç işlemenin bir mahkumun cinsiyeti veya ırkı ile istatistiksel olarak bağlantılı olduğunu tespit ederse - bu, yeniden programlanması gereken cinsiyetçi, ırkçı bir soygun olarak sayılır mı? Nitekim, 2016 yılında yapılan bir araştırma, Amerika Birleşik Devletleri'nde kullanılan tekrar suçlama-tahmin yazılımının Afrikalı Amerikalılara karşı önyargılı olduğunu ve

haksız cezaya katkıda bulundu. [36] Bunlar, yapay zekanın faydalı kalmasını sağlamak için hepimizin üzerinde düşünmesi ve tartışması gereken önemli sorular. Robojudges ile ilgili ya hep ya hiç kararıyla karşı karşıya değiliz, daha ziyade AI'yı yasal sistemimize yerleştirmek istediğimiz kapsam ve hız hakkında bir kararla karşı karşıyayız. İnsan yargıçların tıpkı yarının tıp doktorları gibi yapay zeka tabanlı karar destek sistemlerine sahip olmasını istiyor muyuz? Daha ileri gitmek ve insan yargıçlara itiraz edilebilecek robojudge kararları almak istiyor muyuz, yoksa sonuna kadar gitmek ve ölüm cezaları için bile makinelere son sözü vermek istiyor muyuz?

Yasal Tartışmalar

Şimdiye kadar sadece *uygulama* hukukun; şimdi ona dönelim *içerik.*

Teknolojimize ayak uydurmak için yasalarımızın değişmesi gerektiği konusunda geniş bir fikir birliği var. Örneğin, yukarıda bahsedilen ILOVEYOU solucanını yaratan ve milyarlarca dolar hasara neden olan iki programcı tüm suçlamalardan beraat etti ve o sırada Filipinler'de kötü amaçlı yazılım oluşturmaya karşı herhangi bir yasa olmadığı için serbest kaldı. Teknolojik ilerlemenin hızı hızlanıyor gibi göründüğünden, yasaların daha hızlı güncellenmesi ve geride kalma eğilimi göstermesi gerekiyor. Daha fazla teknoloji meraklısı insanları hukuk okullarına ve hükümetlere sokmak muhtemelen toplum için akıllıca bir harekettir. Ancak seçmenler ve yasa koyucular için yapay zeka tabanlı karar destek sistemleri ve ardından doğrudan robo-yasa koyucular gelmeli mi?

AI ilerlemesini yansıtmak için yasalarımızı en iyi şekilde nasıl değiştirebileceğimiz, büyüleyici bir şekilde tartışmalı bir konudur. Bir anlaşmazlık, mahremiyet ile bilgi özgürlüğü arasındaki gerilimi yansıtıyor. Freedom hayranları, ne kadar az mahremiyetimiz olursa, mahkemelerin sahip olacağı daha fazla kanıt ve kararların o kadar adil olacağını savunuyorlar. Örneğin, hükümet nerede olduklarını ve ne yazdıklarını, tıkladıklarını, söylediklerini ve yaptıklarını kaydetmek için herkesin elektronik cihazlarını kullanırsa, birçok suç kolayca çözülebilir ve başka suçlar önlenebilir. Mahremiyet savunucuları, Orwellci bir gözetim devleti istemediklerini ve öyle olsalar bile, epik boyutlarda totaliter bir diktatörlüğe dönüşme riski olduğunu savunuyorlar. Dahası, makine öğrenme teknikleri, bir kişinin ne düşündüğünü belirlemek için fMRI tarayıcılarından alınan beyin verilerini analiz etmede daha iyi hale geldi.

özellikle doğru mu yoksa yalan mı söylediği hakkında. [37] Yapay zeka destekli beyin tarama teknolojisi mahkeme salonlarında yaygın hale gelirse, bir davanın gerçeklerini tespit etmek için şu anda yorucu olan süreç dramatik bir şekilde basitleştirilebilir ve hızlandırılabilir, bu da daha hızlı yargılamalar ve daha adil kararlar sağlar. Ancak gizlilik savunucuları, bu tür sistemlerin zaman zaman hata yapıp yapmayacağı ve daha temelde, zihinlerimizin hükümetin gözetlemesine kapalı olup olmayacağı konusunda endişelenebilir. Düşünce özgürlüğünü desteklemeyen hükümetler, bu tür teknolojiyi belirli inanç ve görüşlerin sahiplenilmesini suç haline getirmek için kullanabilir. Nerede olur *sen* adalet ve mahremiyet arasındaki ve toplumu korumak ile kişisel özgürlüğü korumak arasındaki çizgiyi mi çiziyor? Onu nereye çizerseniz çizin, yavaş yavaş ama amansızca, bunu telafi etmek için mahremiyetin azaltılmasına doğru hareket edecek mi?

kanıt sahteciliği yapmak daha kolay Örneğin, yapay zeka sizin suç işlediğinize dair tamamen gerçekçi sahte videolar oluşturabildiğinde, hükümetin herkesin nerede olduğunu her zaman izlediği ve gerekirse size sağlam bir mazeret sağlayabileceği bir sisteme oy verecek misiniz?

Bir başka büyüleyici tartışma, AI araştırmasının düzenlenip düzenlenmeyeceği veya daha genel olarak, yararlı bir sonucun olasılığını en üst düzeye çıkarmak için politika yapıcıların AI araştırmacılarına hangi teşvikleri vermesi gerektiğidir. Bazı yapay zeka araştırmacıları, acil ihtiyaç duyulan yeniliği (örneğin, hayat kurtaran otonom arabalar) gereksiz yere geciktireceklerini ve yeraltında ve / veya diğer ülkelerde en son yapay zeka araştırmalarını sürdüreceklerini iddia ederek, yapay zeka geliştirmenin her türlü düzenlemesine karşı çıktılar. izin veren hükümetler. İlk bölümde bahsi geçen Porto Riko Yararlı Yapay Zeka konferansında Elon Musk, şu anda hükümetlerden ihtiyacımız olan şeyin gözetim değil, içgörü olduğunu savundu: özellikle, YZ'nin ilerlemesini izleyebilecek ve gerektiğinde yönlendirebilecek hükümet pozisyonlarındaki teknik olarak yetenekli insanlar. yolun aşağısında. Ayrıca, hükümet düzenlemelerinin bazen ilerlemeyi bastırmaktan ziyade besleyebileceğini savundu: örneğin, kendi kendine giden arabalara yönelik hükümet güvenlik standartları, kendi kendine giden araba kazalarının sayısını azaltmaya yardımcı olabilirse, o zaman halkın tepkisi daha az olasıdır ve yeni teknoloji hızlandırılabilir. Bu nedenle, güvenlik konusunda en bilinçli AI şirketleri, daha az titiz rakipleri yüksek güvenlik standartlarına uymaya zorlayan düzenlemeyi tercih edebilir.

Yine bir başka ilginç yasal tartışma, makinelere hakların verilmesidir. Otonom arabalar ABD'deki yıllık 32.000 ölüm oranını yarı yarıya düşürdüyse, belki de otomobil üreticileri 16.000 teşekkür notu değil, 16.000 dava alacak. Öyleyse, sürücüsüz bir araba bir kazaya neden olursa, bundan kim sorumlu olmalıdır - yolcuları mı, sahibi mi yoksa üreticisi mi? Hukuk bilgini David Vladeck dördüncü bir cevap önerdi: arabanın kendisi! Spesifik olarak, sürücüsüz arabaların araba sigortası yaptırmasına izin verilmesini (ve gerekli) önermektedir. Bu şekilde, sterlin güvenlik siciline sahip modeller, çok düşük, muhtemelen insan sürücüler için mevcut olandan daha düşük primler için uygun olurken, özensiz üreticilerin kötü tasarlanmış modelleri yalnızca onları sahip olmak için aşırı derecede pahalı hale getiren sigorta poliçelerine hak kazanacaktır.

Ancak, arabalar gibi makinelerin sigorta poliçelerine sahip olmasına izin veriliyorsa, onlar da para ve mülk sahibi olabilmeli mi? Öyleyse, akıllı bilgisayarların borsada para kazanmasını ve onu çevrimiçi hizmetler satın almak için kullanmasını yasal olarak engelleyen hiçbir şey yoktur. Bir bilgisayar, çalışması için insanlara ödeme yapmaya başladığında, insanların yapabileceği her şeyi başarabilir. AI sistemleri sonunda daha iyi hale gelirse

Yatırım yapan insanlardan (ki zaten bazı alanlarda zaten varlar), bu, ekonomimizin çoğunun sahip olduğu ve makineler tarafından kontrol edildiği bir duruma yol açabilir. Bu bizim istediğimiz şey mi? Kulağa çok uzak geliyorsa, ekonomimizin çoğunun hâlihazırda başka bir insan olmayan varlığa ait olduğunu düşünün: Genellikle içlerindeki herhangi bir kişiden daha güçlü olan ve bir dereceye kadar kendi hayatlarını sürdürebilen şirketler.

Makinelere mülk sahibi olma hakları vermekte sorun yoksa, onlara oy kullanma hakkı vermeye ne dersiniz? Eğer öyleyse, yeterince zenginse bulutta kendisinin trilyonlarca kopyasını önemsiz bir şekilde yapabilmesine ve böylece tüm seçimlere karar vereceğini garanti etmesine rağmen, her bilgisayar programı bir oy almalı mı? Değilse, insan zihnine göre makine zihinlerine karşı hangi ahlaki temelde ayrımcılık yapıyoruz? Bizim gibi öznel bir deneyime sahip olma anlamında makine zihinlerinin bilinçli olması bir fark yaratır mı? Bir sonraki bölümde dünyamızın bilgisayar kontrolü ile ilgili bu tartışmalı soruları ve 8. bölümde makine bilinciyle ilgili soruları daha derinlemesine inceleyeceğiz.

Silahlar

Çok eski zamanlardan beri insanlık kıtlık, hastalık ve savaştan acı çekti. Yapay zekanın kıtlığı ve hastalığı azaltmaya nasıl yardımcı olabileceğinden daha önce bahsetmiştik, peki ya savaş? Bazıları nükleer silahların kendilerine sahip olan ülkeler arasındaki savaşı caydırdığını, çünkü çok korkunç olduklarını iddia ediyorlar, peki bütün ülkelerin savaşı sonsuza dek sona erdirme umuduyla daha da korkunç AI tabanlı silahlar yapmasına izin vermeye ne dersiniz? Bu argümanla ikna olmuyorsanız ve gelecekteki savaşların kaçınılmaz olduğuna inanıyorsanız, bu savaşları daha insani hale getirmek için yapay zekayı kullanmaya ne dersiniz? Savaşlar yalnızca makinelerle savaşan makinelerden oluşuyorsa, o zaman hiçbir insan askerin veya sivilin öldürülmesine gerek yoktur. Dahası, gelecekteki yapay zeka destekli insansız hava araçları ve diğer otonom silah sistemleri (AWS; rakipleri tarafından "katil robotlar" olarak da bilinir) umarız insan askerlerden daha adil ve rasyonel hale getirilebilir:



Şekil 3.4: Bugünün askeri insansız hava araçları (bu ABD Hava Kuvvetleri MQ-1 Predator gibi) insanlar tarafından uzaktan kontrol edilirken, gelecekteki yapay zeka destekli insansız hava araçları, kimi hedef alacaklarına karar vermek için bir algoritma kullanarak insanları döngüden çıkarma potansiyeline sahiptir. ve öldür.

Döngüdeki AHuman

Peki ya otomatik sistemler hatalı, kafa karıştırıcıysa veya beklendiği gibi davranmıyorsa? Aegis sınıfı kruvazörlere yönelik US Phalanx sistemi, gemi karşıtı füzeler ve uçaklar gibi tehditleri otomatik olarak algılar, izler ve saldırır. USS *Vincennes* Aegis sistemine atıfta bulunularak Robocruiser lakaplı güdümlü bir füze kruvazörü idi ve 3 Temmuz 1988'de İran-Irak savaşı sırasında İran savaş gemileriyle bir çatışmanın ortasında, radar sistemi gelen bir uçak konusunda uyardı. Kaptan William Rodgers III, dalış yapan bir İran F-14 savaş uçağı tarafından saldırıya uğradıklarını belirterek Aegis sistemine ateş etme onayı verdi. O sırada fark etmediği şey, sivil bir İran yolcu uçağı olan Iran Air Flight 655'i düşürdükleri, gemideki 290 kişiyi öldürdükleri ve uluslararası öfkeye neden olduklarıydı. Sonraki soruşturma, radar ekranındaki hangi noktaların sivil uçaklar olduğunu (Uçuş 655, normal günlük uçuş yolunu takip etti ve sivil uçak transponderini açtı) veya hangi noktaların alçaldığını (bir saldırı için olduğu gibi) otomatik olarak göstermeyen kafa karıştırıcı bir kullanıcı arayüzünü ortaya çıkardı. vs. yükselen (655 sefer sayılı uçağın Tahran'dan kalktıktan sonra yaptığı gibi). Bunun yerine, otomatik sistem gizemli uçakla ilgili bilgi için sorgulandığında, "alçalıyor" olarak rapor edildi çünkü bu, donanma tarafından uçakları izlemek için kullanılan bir numarayı kafa karıştırıcı bir şekilde yeniden tahsis ettiği farklı bir uçağın durumuydu: alçalan bunun yerine Umman Körfezi'nde çok uzakta faaliyet gösteren bir ABD yüzey muharebe hava devriye uçağı.

Bu örnekte, döngüde son kararı veren ve zaman baskısı altında otomatik sistemin kendisine söylediklerine çok fazla güvenen bir insan vardı. Şimdiye kadar, dünyanın dört bir yanındaki savunma yetkililerine göre, tüm konuşlandırılmış silah sistemlerinde, kara mayınları gibi düşük teknolojili bubi tuzakları hariç, döngüde bir insan var. Ancak hedefleri tamamen kendi başlarına seçen ve onlara saldıran gerçekten otonom silahların geliştirilmesi şu anda devam ediyor. Hız kazanmak için tüm insanları döngüden çıkarmak askeri açıdan cazip geliyor: anında tepki verebilen tamamen otonom bir drone ile daha yavaş tepki veren bir drone arasındaki it dalaşında, çünkü dünyanın öbür ucundaki bir insan tarafından uzaktan kumanda ediliyor. düşünmek kazanır mı?

Ancak, döngüde bir insan olduğu için son derece şanslı olduğumuz yakın çağrılar oldu. 27 Ekim 1962'de Küba Füze Krizi sırasında, on bir ABD Donanması muhrip ve uçak gemisi USS *Randolph* vardı

Sovyet denizaltısı B-59'u Küba yakınlarında, ABD “karantina” bölgesinin dışındaki uluslararası sularda köşeye sıkıştırdı. Bilmedikleri şey, denizaltının pillerinin bitmesi ve klimanın durması nedeniyle gemideki sıcaklığın 45 ° C'yi (113 ° F) geçtiğiydi. Karbondioksit zehirlenmesinin eşiğinde, birçok mürettebat üyesi bayılmıştı. Mürettebatın günlerdir Moskova ile hiçbir bağlantısı yoktu ve III.Dünya Savaşı'nın çoktan başlayıp başlamadığını bilmiyordu. Sonra Amerikalılar, mürettebatın haberi olmadan Moskova'ya denizaltıyı yüzeye çıkmaya zorlamak olduğunu söylediler. Mürettebat üyesi VP Orlov, "Sonun bu olduğunu düşündük," diye hatırladı. “Sanki birilerinin sürekli balyozla patlattığı metal bir fıçıda oturuyormuşsunuz gibi hissettim. Amerikalıların da bilmediği şey, B-59 mürettebatının Moskova ile temizlemeden fırlatmaya yetkilendirildikleri bir nükleer torpidoya sahip olduğuydu. Nitekim Kaptan Savitski nükleer torpidoyu fırlatmaya karar verdi. Torpido subayı Valentin Grigorievich, "Öleceğiz, ama hepsini batıracağız - donanmamızı utandırmayacağız!" Neyse ki, fırlatma kararının gemideki üç subay tarafından onaylanması gerekiyordu ve içlerinden biri, Vasili Arkhipov hayır dedi. Kararı III.Dünya Savaşı'nın önüne geçmiş olsa da, Arkhipov'u çok az kişinin duymuş olması çok üzücü. ama hepsini batıracağız - donanmamızı küçük düşürmeyeceğiz! " Neyse ki, fırlatma kararının gemideki üç subay tarafından onaylanması gerekiyordu ve içlerinden biri, Vasili Arkhipov hayır dedi. Kararı III.Dünya Savaşı'nın önüne geçmiş olsa da, Arkhipov'u çok az kişinin duymuş olması çok üzücü. ama hepsini batıracağız - donanmamızı küçük düşürmeyeceğiz! " Neyse ki, fırlatma kararının gemideki üç subay tarafından onaylanması gerekiyordu ve içlerinden biri, Vasili Arkhipov hayır dedi. Kararı III.Dünya Savaşı'nın önüne geçmiş olsa da, Arkhipov'u çok az kişinin duymuş olması çok üzücü.

modern tarihte insanlığa en değerli katkı. [38] Ayrıca, B-59 döngüsünde insan bulunmayan otonom yapay zeka kontrollü bir denizaltı olsaydı neler olabileceğini düşünmek de ayıptır.

Yirmi yıl sonra, 9 Eylül 1983'te, süper güçler arasındaki gerginlikler yeniden yükseldi: Sovyetler Birliği, yakın zamanda "kötü bir imparatorluk" olarak adlandırıldı

ABD başkanı Ronald Reagan ve daha geçen hafta, hava sahasına giren bir Korean Airlines yolcu uçağını düşürerek 269 kişiyi öldürmüştü.

- bir ABD kongre üyesi dahil. Şimdi otomatik bir Sovyet erken uyarı sistemi, Amerika Birleşik Devletleri'nin Sovyetler Birliği'ne beş adet kara tabanlı nükleer füze fırlattığını ve bunun yanlış bir alarm olup olmadığına karar vermek için Subay Stanislav Petrov'a yalnızca birkaç dakika kaldığını bildirdi. Uydunun düzgün çalıştığı tespit edildi, bu nedenle aşağıdaki protokol, onu gelen bir nükleer saldırıyı bildirmeye yönlendirirdi. Bunun yerine, Amerika Birleşik Devletleri'nin yalnızca beş füzeyle saldırmasının olası olmadığını düşünerek içgüdüsüne güvendi ve komutanlarına bunun doğru olduğunu bilmeden yanlış bir alarm olduğunu bildirdi. Daha sonra, bir uydunun, Güneş'in bulutların tepelerinden yansımalarını,

roket motorları. [39] Petrov yerine uygun protokolü uygun şekilde takip eden bir AI sistemi alsaydı ne olurdu merak ediyorum.

Bir Sonraki Silahlanma Yarışı?

Şimdiye kadar hiç şüphesiz tahmin ettiğiniz gibi, şahsen otonom silah sistemleri hakkında ciddi endişelerim var. Ama size asıl endişemden bahsetmeye bile başlamadım: AI silahlarındaki silahlanma yarışının son noktası. Temmuz 2015'te bu endişeyi Stuart Russell ile birlikte aşağıdaki açık mektupta dile getirdim:

Future of Life Enstitüsü'ndeki meslektaşlarımdan yararlı geri bildirimlerle: [40]

ÖZERK SİLAHLAR:

Yapay Zeka ve Robotik Araştırmacılarından Açık Mektup

Otonom silahlar, insan müdahalesi olmadan hedefleri seçer ve bunlara saldırır. Örneğin, önceden tanımlanmış belirli kriterleri karşılayan kişileri arayabilen ve ortadan kaldırabilen, ancak seyir füzelerini veya insanların tüm hedefleme kararlarını aldığı uzaktan kumandalı insansız hava araçlarını içermeyen silahlı quadcopter'lar içerebilirler. Yapay Zeka (AI) teknolojisi, bu tür sistemlerin konuşlandırılmasının yasal olarak on yıllar içinde değil, yıllar içinde yasal olarak mümkün değilse de pratikte mümkün olduğu ve risklerin yüksek olduğu bir noktaya ulaştı: otonom silahlar, barut ve nükleer silahlardan sonra savaşta üçüncü devrim olarak tanımlandı. silâh.

Otonom silahlar lehine ve aleyhine pek çok tartışma yapılmıştır, örneğin insan askerleri makinelerle değiştirmenin mal sahibi için kayıpları azaltarak iyi, ancak savaşa gitme eşiğini düşürerek kötü olduğu. Bugün insanlık için temel soru, küresel bir AI silahlanma yarışına başlamak mı yoksa başlamasını engellemek mi? Herhangi bir büyük askeri güç, AI silah geliştirme ile ilerlerse, küresel bir silahlanma yarışı neredeyse kaçınılmazdır ve bu teknolojik yörüngenin son noktası açıktır: otonom silahlar, yarının Kalaşnikofları olacak. Nükleer silahlardan farklı olarak, maliyetli veya elde edilmesi zor hammaddelere ihtiyaç duymazlar, bu nedenle her yerde bulunur ve tüm önemli askeri güçlerin toplu üretmesi için ucuz hale gelirler. Karaborsaya çıkmaları ve teröristlerin eline geçmeleri an meselesi olacak. Halklarını daha iyi kontrol etmek isteyen diktatörler, etnik temizlik yapmak isteyen savaş ağaları, vb. Otonom silahlar, suikastlar, ulusların istikrarını bozma, nüfusları bastırma ve belirli bir etnik grubu seçerek öldürme gibi görevler için idealdir. Bu nedenle, askeri bir AI silahlanma yarışının insanlık için faydalı olmayacağına inanıyoruz. Yapay zekanın savaş alanlarını insanlar için daha güvenli hale getirmesinin birçok yolu vardır, özellikle

siviller, insanları öldürmek için yeni araçlar yaratmadan.

Tıpkı çoğu kimyager ve biyologun kimyasal veya biyolojik silahlar yapmakla ilgilenmemesi gibi, çoğu AI araştırmacısının AI silahları inşa etmekle hiç ilgisi yoktur ve başkalarının bunu yaparak kendi alanlarını lekelemesini istemezler, bu da potansiyel olarak AI'ya karşı, kendi gelecekteki toplumsal faydalar. Aslında kimyagerler ve biyologlar, tıpkı çoğu fizikçinin uzay temelli nükleer silahları ve kör edici lazer silahlarını yasaklayan anlaşmaları desteklemesi gibi, kimyasal ve biyolojik silahları başarıyla yasaklayan uluslararası anlaşmaları geniş ölçüde desteklediler.

Endişelerimizi yalnızca pasifist ağaç kucaklayıcılarından geldiği için reddetmeyi zorlaştırmak için, mektubumuzu olabildiğince çok sayıda hardcore AI araştırmacısı ve robotikçi tarafından imzalatmak istedim. Uluslararası Robotik Silah Kontrolü Kampanyası, daha önce katil robotların yasaklanması çağrısında bulunan yüzlerce imzacı topladı ve daha da iyisini yapabileceğimizden şüpheleniyordum. Meslek kuruluşlarının siyasi olarak yorumlanabilecek bir amaç için devasa üye e-posta listelerini paylaşmaktan çekineceklerini biliyordum, bu yüzden araştırmacıların isim ve kurum listelerini çevrimiçi belgelerden bir araya getirdim ve e-posta adreslerini MTurk'te bulma görevinin reklamını yaptım. —Amazon Mechanical Turk kitle kaynak platformu. Çoğu araştırmacının e-posta adresleri üniversite web sitelerinde listelenir ve yirmi dört saat ve 54 dolar sonra, Yapay Zekayı Geliştirme Derneği'nin (AAAI) Üyeleri seçilmek için yeterince başarılı olan yüzlerce yapay zeka araştırmacısından oluşan bir posta listesinin gururlu sahibiydim. Bunlardan biri, listedeki herkese e-posta göndermeyi ve kampanyamıza öncülük etmeye yardımcı olmayı nazikçe kabul eden İngiliz-Avustralyalı yapay zeka profesörü Toby Walsh'du. Dünyanın dört bir yanındaki MTurk çalışanları, Toby için yorulmadan ek posta listeleri üretti ve çok geçmeden, Google, Facebook, Microsoft ve Tesla'dan altı geçmiş AAAI başkanı ve AI endüstri liderleri de dahil olmak üzere 3.000'den fazla AI araştırmacısı ve robotik araştırmacısı açık mektubumuzu imzaladılar. FLI gönüllülerinden oluşan bir ordu imza listelerini doğrulamak için yorulmadan çalıştı ve Bill Clinton ve Sarah Connor gibi sahte girişleri kaldırdı. Stephen Hawking dahil 17.000'den fazla kişi de imzaladı,

Biyologlar ve kimyagerler bir zamanlar tavır aldıklarından, alanları artık biliniyor

temel olarak biyolojik ve kimyasal silahlardan ziyade faydalı ilaç ve malzemeler yaratmak için. Yapay zeka ve robotik toplulukları şimdi de konuşmuştu: Mektup imzacıları, tarlalarının insanları öldürmek için yeni yollar yaratmak için değil, daha iyi bir gelecek yaratmasıyla da bilinmesini istiyorlardı. Ancak yapay zekanın gelecekteki ana kullanımı sivil mi yoksa askeri mi olacak? Bu bölümde birincisine daha fazla sayfa harcamış olsak da, yakın zamanda ikinciye daha fazla para harcıyor olabiliriz - özellikle de askeri bir AI silahlanma yarışı başlarsa. Sivil yapay zeka yatırım taahhütleri 2016'da bir milyar doları aştı, ancak bu, Pentagon'un yapay zeka ile ilgili projeler için 2017 mali yılı bütçe talebinin 12-15 milyar dolar olması nedeniyle cüce kaldı ve Çin ve Rusya, Savunma Bakan Yardımcısı Robert Work'ün söylediklerini muhtemelen dikkate alacaklar. bu duyurulduğunda:

perde." [41]

Uluslararası Bir Antlaşma Olmalı mı?

Şu anda bir çeşit katil robot yasağını müzakere etmeye yönelik büyük bir uluslararası itici güç olsa da, ne olacağı hala belirsiz ve eğer varsa, ne olacağı konusunda devam eden canlı bir tartışma var. *meli* olmak. Önde gelen pek çok paydaş, dünya güçlerinin AWS araştırmalarına ve kullanımına rehberlik edecek bir tür uluslararası düzenlemeler hazırlaması gerektiği konusunda hemfikir olsa da, tam olarak neyin yasaklanması gerektiği ve bir yasağın nasıl uygulanacağı konusunda daha az fikir birliği var. Örneğin, yalnızca ölümcül otonom silahlar yasaklanmalı mı, yoksa insanları ciddi şekilde yaralayanlar, örneğin onları kör ederek mi yasaklanmalı? Geliştirmeyi, üretimi veya mülkiyeti yasaklayacak mıydık? Yasak, tüm otonom silah sistemlerine mi uygulanmalı yoksa mektubumuzun da belirttiği gibi, otonom uçaksavar silahları ve füze savunmaları gibi savunma sistemlerine izin veren yalnızca saldırgan olanlar mı? İkinci durumda, AWS, düşman bölgesine taşınması kolay olsa bile savunma olarak kabul edilmeli mi? Otonom bir silahın çoğu bileşeninin de çifte sivil kullanıma sahip olduğu göz önüne alındığında, bir anlaşmayı nasıl uygulayabilirsiniz? Örneğin,

Bazı tartışmacılar, etkili bir AWS anlaşması tasarlamanın umutsuzca zor olduğunu ve bu nedenle denemememiz gerektiğini savundu. Öte yandan, John F. Kennedy, Ay görevlerini duyururken, başarının insanlığın geleceğine büyük fayda sağlayacağı zaman zor şeylerin denemeye değer olduğunu vurguladı. Dahası, birçok uzman, biyolojik ve kimyasal silahlara yönelik yasakların, yaptırımın zor olduğu kanıtlanmasına rağmen, önemli hile ile birlikte değerli olduğunu, çünkü yasakların kullanımlarını sınırlayan ciddi damgalamaya neden olduğunu savunuyor.

Henry Kissinger ile 2016'da bir akşam yemeğinde tanıştım ve ona biyolojik silah yasağındaki rolü hakkında soru sorma fırsatı buldum. ABD ulusal güvenlik danışmanıyken, Başkan Nixon'u bir yasağın ABD ulusal güvenliği için iyi olacağına nasıl ikna ettiğini açıkladı. Doksan iki yaşındaki bir çocuk için zihninin ve hafızasının ne kadar keskin olmasından etkilendim ve içeriden bakış açısını duymak beni büyüledi. Amerika Birleşik Devletleri, konvansiyonel ve nükleer güçleri sayesinde zaten süper güç statüsüne sahip olduğundan, sonuçları belirsiz olan dünya çapında bir biyolojik silah silahlanma yarışından kazanmaktan çok kaybedecek daha çok şey vardı. Başka bir deyişle, zaten en iyiyseniz, "Kırılmamışsa, düzeltmeyin" özdeyişini takip etmek mantıklıdır. Stuart Russell yemek sonrası sohbetimize katıldı ve aynı tartışmanın nasıl yapılabileceğini tartıştık.

ölümcül otonom silahlar hakkında: Bir silahlanma yarışından en çok kazanmayı bekleyenler süper güçler değil, küçük haydut devletler ve geliştirildikten sonra karaborsa yoluyla silahlara erişim sağlayan teröristler gibi devlet dışı aktörler.

Seri olarak üretildikten sonra, küçük yapay zeka destekli katil dronların bir akıllı telefondan biraz daha pahalı olması muhtemeldir. İster bir politikacıya suikast düzenlemek isteyen bir terörist, isterse eski kız arkadaşından intikam almak isteyen reddedilmiş bir aşık olsun, tek yapmaları gereken hedefin fotoğrafını ve adresini katil drona yüklemek: daha sonra hedefe uçabilir, tespit edip ortadan kaldırabilir. Kimsenin sorumlu olduğunu bilmemesini sağlamak için kişi ve kendini yok etme. Alternatif olarak, etnik temizliğe meyilli olanlar için, yalnızca belirli bir ten rengine veya etnik kökene sahip kişileri öldürmek için kolayca programlanabilir. Stuart, bu tür silahlar ne kadar akıllı olursa, öldürme başına o kadar az malzeme, ateş gücü ve paraya ihtiyaç duyulacağını düşünüyor. Örneğin, insanları gözlerine vurarak minimum patlayıcı güç kullanarak ucuza öldüren yaban arısı büyüklüğündeki dronlardan korkuyor, Bu, küçük bir merminin bile beyne girmesine izin verecek kadar yumuşaktır. Ya da metal pençelerle kafasına takılıp küçük şekilli bir yük ile kafatasına girebilirler. Bu türden bir milyon insansız hava aracı, tek bir kamyonun arkasından gönderilebilirse, o zaman kişi yepyeni bir türden korkunç bir kitle imha silahına sahip olur: yalnızca belirli bir kategorideki insanı seçerek öldürebilen, herkesi ve diğer her şeyi zarar görmeyen bir silah. .

Ortak bir karşı argüman, katil robotları etik hale getirerek bu tür endişeleri ortadan kaldırabileceğimizdir - örneğin, böylece onlar yalnızca düşman askerlerini öldürürler. Ancak bir yasağı uygulama konusunda endişeleniyorsak, o zaman düşmanın otonom silahlarının% 100 etik olması şartını uygulamak, ilk etapta üretilmemelerini zorunlu kılmaktan nasıl daha kolay olabilir? Ve uygar ulusların iyi eğitimli askerlerinin, robotların daha iyi yapabileceği savaş kurallarına uymada o kadar kötü olduğu ve aynı zamanda haydut ulusların, diktatörlerin ve terörist grupların kurallara uymada çok başarılı olduğu iddia edilebilir mi? Robotları asla bu kuralları ihlal edecek şekilde konuşlandırmayı seçmeyecekleri bir savaş mı?

Sanal savaş

Yapay zekanın bir başka ilginç askeri yönü de, siber savaş yoluyla kendi silahlarınızı oluşturmadan bile düşmanınıza saldırmanıza izin verebilmesidir. Geleceğin getireceği şeylerin küçük bir başlangıcı olarak, yaygın olarak ABD ve İsrail hükümetlerine atfedilen Stuxnet solucanı, İran'ın nükleer zenginleştirme programındaki hızlı dönen santrifüjlere bulaştı ve kendilerini parçalamalarına neden oldu. Toplum ne kadar otomatikleşirse ve saldıran yapay zeka ne kadar güçlü olursa, siber savaş da o kadar yıkıcı olabilir. Düşmanınızın sürücüsüz arabalarını, otomatik pilotlu uçaklarını, nükleer reaktörleri, endüstriyel robotları, iletişim sistemlerini, finansal sistemlerini ve güç şebekelerini hackleyip çarpabilirseniz, ekonomisini etkili bir şekilde çökertebilir ve savunmasını zayıflatabilirsiniz. Bazı silah sistemlerini de hackleyebilirseniz, daha da iyisi.

Bu bölüme, yapay zekanın insanlığa fayda sağlaması için yakın vadeli fırsatların ne kadar muhteşem olduğunu araştırarak başladık - eğer onu sağlam ve hacklenemez hale getirmeyi başarırsak. Yapay zeka, yapay zeka sistemlerini daha sağlam hale getirmek için kullanılabilse de, böylece siber savaş savunmasına yardımcı olabilir, ancak yapay zeka açıkça saldırıya da yardımcı olabilir. Savunmanın galip gelmesini sağlamak, AI geliştirme için en önemli kısa vadeli hedeflerden biri olmalıdır - aksi takdirde oluşturduğumuz tüm harika teknolojiler aleyhimize dönebilir!

İşler ve Ücretler

Bu bölümde şu ana kadar, yapay zekanın bizi nasıl etkileyeceğine odaklandık. *tüketiciler* uygun fiyatlarla dönüştürücü yeni ürün ve hizmetleri mümkün kılarak. Ama bizi nasıl etkileyecek *işçiler* iş piyasasını dönüştürerek mi? İnsanları gelirden veya amaçtan yoksun bırakmadan otomasyon yoluyla refahımızı nasıl artıracağımızı bulabilirsek, o zaman isteyen herkes için boş zaman ve benzeri görülmemiş bir zenginlikle harika bir gelecek yaratma potansiyeline sahibiz. MIT'deki meslektaşlarımdan biri olan ekonomist Erik Brynjolfsson'dan çok az insan bu konuda daha uzun süre ve daha fazla düşündü. Her zaman bakımlı ve kusursuz giyinmiş olmasına rağmen, İzlanda mirasına sahip ve bazen işletme okulumuzda uyum sağlamak için vahşi kırmızı bir Viking sakalını ve yelesini daha yeni kestiğini hayal edemiyorum. Kesinlikle çılgın fikirlerini geri çekmedi ve iyimser iş piyasası vizyonuna "Dijital Atina" diyor. Antik Atina vatandaşlarının demokrasinin tadını çıkarabilecekleri boş zamanlarında yaşamalarının nedeni, sanat ve oyunlar esas olarak işin çoğunu yapacak kölelere sahip olmalarıdır. Ama neden köleleri yapay zeka destekli robotlarla değiştirerek herkesin keyif alabileceği dijital bir ütopya yaratmayasınız? Erik'in yapay zeka güdümlü ekonomisi yalnızca stresi ve angarya işini ortadan kaldırmak ve bugün istediğimiz her şeyin bolluğunu üretmekle kalmayacak, aynı zamanda günümüz tüketicilerinin henüz istediklerini fark etmedikleri harika yeni ürünler ve hizmetler sunacaktı.

Teknoloji ve Eşitsizlik

Herkesin saatlik maaşı her geçen yıl artmaya devam ederse, bugün bulunduğumuz yerden Erik'in Dijital Atina'sına gidebiliriz, böylece daha fazla eğlence isteyenler, yaşam standartlarını iyileştirmeye devam ederken kademeli olarak daha az çalışabilirler.

Şekil 3.5 İkinci Dünya Savaşı'ndan 1970'lerin ortalarına kadar Amerika Birleşik Devletleri'nde olan şeyin tam olarak bu olduğunu gösteriyor: gelir eşitsizliği olmasına rağmen, pastanın toplam boyutu neredeyse herkesin daha büyük bir dilim alacağı şekilde büyüdü. Ama sonra, Erik'in ilk itiraf ettiği gibi, bir şeyler değişti: şekil 3.5 Ekonomi büyümeye ve ortalama geliri artırmaya devam etmesine rağmen, son kırk yılda elde edilen kazanımların en zenginlere, çoğunlukla en tepedeki% 1'e gittiğini, en yoksul% 90'ın ise gelirlerinin durgun olduğunu gördüğünü gösteriyor. Gelire değil servete bakarsak, eşitsizlikte ortaya çıkan büyüme daha da belirgindir. En alt% 90'ı için

ABD hanehalkları, ortalama net değer 2012'de yaklaşık 85.000 dolardı - yirmi beş yıl önceki ile aynı - en tepedeki% 1 enflasyonunu ikiye katladı-

o dönemde serveti 14 milyon dolara ayarladı. [42] 2013'te dünya nüfusunun en alt yarısının (3,6 milyardan fazla insan) toplam servetinin, dünya nüfusununkiyle aynı olduğu uluslararası düzeyde farklılıklar daha da aşırıdır.

dünyanın en zengin sekiz insanı [43] - En üstteki zenginlik kadar en alttaki yoksulluk ve kırılganlığı vurgulayan bir istatistik. 2015 Porto Riko konferansımızda Erik, toplanan yapay zeka araştırmacılarına, yapay zeka ve otomasyondaki ilerlemenin ekonomik pastayı büyütmeye devam edeceğini düşündüğünü, ancak herkesin, hatta çoğu insanın yararlanacağı bir ekonomik yasa olmadığını söyledi.

Ekonomistler arasında eşitsizliğin arttığı konusunda geniş bir fikir birliği olsa da, eğilimin neden ve devam edip etmeyeceği konusunda ilginç bir tartışma var. Siyasi yelpazenin sol tarafındaki tartışmacılar genellikle ana nedenin küreselleşme ve / veya zenginler için vergi indirimleri gibi ekonomik politikalar olduğunu savunuyorlar. Ancak Erik Brynjolfsson ve MIT'deki iş arkadaşı Andrew McAfee,

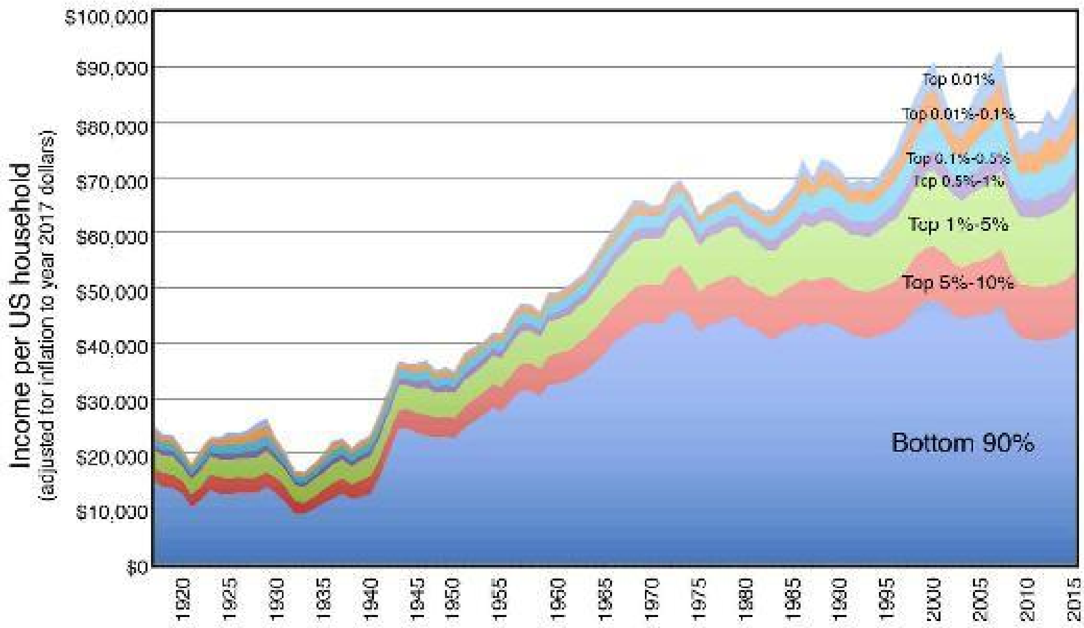
neden başka bir şeydir: teknoloji. [44] Spesifik olarak, dijital teknolojinin eşitsizliği üç farklı yoldan yönlendirdiğini savunuyorlar.

Birincisi, eski işleri daha fazla beceri gerektiren işlerle değiştirerek, teknoloji eğitimli kişileri ödüllendirdi: 1970'lerin ortalarından bu yana, ortalama lise terk% 30 maaş alırken, yüksek lisans derecesine sahip olanlar için maaşlar yaklaşık% 25 arttı

kesmek. [45]

İkincisi, 2000 yılından bu yana, şirket gelirlerinin giderek daha büyük bir kısmının, orada çalışanların aksine şirketlere sahip olanlara gittiğini ve otomasyon devam ettiği sürece makinelerin sahiplerini beklememiz gerektiğini iddia ediyorlar. pastanın büyüyen bir kısmını almak için. Sermayenin emek üzerindeki bu sınırı, teknolojik vizyon sahibi Nicholas Negroponte'nin atomlar değil, hareketli parçalar olarak tanımladığı büyüyen dijital ekonomi için özellikle önemli olabilir. Artık kitaplardan filmlere ve vergi hazırlama araçlarına kadar her şey dijital hale geldiğine göre, ek kopyalar dünya çapında ek çalışanlar işe almadan temelde sıfır maliyetle satılabilir. Bu, gelirin çoğunun işçilerden ziyade yatırımcılara gitmesine izin verir ve Detroit'in "Büyük 3" ün (GM,

Market. [47]



Şekil 3.5: Ekonomi geçen yüzyılda ortalama geliri nasıl artırdı ve bu gelirin ne kadarı farklı gruplara gitti. 1970'lerden önce, zenginlerin ve yoksulların hepsinin kilitli bir adımda daha iyi durumda olduğu görülüyordu, ardından kazançların çoğu ilk% 1'e gitti.

En düşük% 90 ortalama olarak sıfıra yakın kazanmıştır. 46 Miktarlar enflasyona göre 2017 yılı dolar olarak düzeltildi.

Üçüncüsü, Erik ve işbirlikçileri, dijital ekonominin çoğu zaman süper starlara diğerlerinden daha fazla fayda sağladığını savunuyor. Harry Potter yazarı JK Rowling, milyarderler kulübüne katılan ilk yazar oldu ve Shakespeare'den çok daha zengin oldu çünkü hikayeleri metin, film ve oyun biçiminde milyarlarca insana çok düşük maliyetle aktarılabiliyordu. Benzer şekilde, Scott Cook, insan vergi hazırlayıcılarından farklı olarak indirilerek satılabilen TurboTax vergi hazırlama yazılımında bir milyar kazandı. Çoğu insan en iyi onuncu vergi hazırlama yazılımı için çok az ödeme yapmaya veya hiçbir şey ödemeye istekli olduğundan, piyasada yalnızca mütevazı sayıda süperstar için yer var. Bu, dünyadaki tüm ebeveynler çocuklarına bir sonraki JK Rowling, Gisele Bündchen, Matt Damon, Cristiano Ronaldo, Oprah Winfrey veya Elon Musk olmalarını tavsiye ederse,

Çocuklar İçin Kariyer Önerileri

Peki ne kariyer tavsiyesi *meli* çocuklarımıza verir miyiz? Benimkini makinelerin şu anda kötü olduğu ve bu nedenle yakın gelecekte otomatik hale gelme ihtimalinin düşük olduğu mesleklere girmeye teşvik ediyorum. Çeşitli işlerin makineler tarafından ne zaman devralınacağına dair son tahminler, bir kariyer hakkında sorulacak birkaç yararlı soruyu belirler
bunun için kendini eğitmeye karar vermeden önce. [48] Örneğin:

- İnsanlarla etkileşime girmeyi ve sosyal zekayı kullanmayı gerektiriyor mu?
- Yaratıcılığı ve akıllı çözümler üretmeyi içeriyor mu?
- Tahmin edilemeyen bir ortamda çalışmayı gerektiriyor mu?

Bu sorulardan ne kadar çok cevabını evet olarak verirseniz, kariyer seçiminiz o kadar iyi olacaktır. Bu, nispeten güvenli bahislerin öğretmen, hemşire, doktor, diş hekimi, bilim adamı, girişimci, programcı, mühendis, avukat, sosyal hizmet uzmanı, din adamı, sanatçı, kuaför veya masaj terapisti olmayı içerdiği anlamına gelir.

Bunun aksine, öngörülebilir bir ortamda yüksek oranda tekrarlayan veya yapılandırılmış eylemler içeren işler, otomatikleştirilmeden önce uzun sürmeyebilir. Bilgisayarlar ve endüstriyel robotlar, bu türden en basit işleri uzun zaman önce devraldı ve teknolojinin geliştirilmesi, tele-pazarlamacılardan depo çalışanlarına, kasiyerlere, tren operatörlerine, fırıncılara ve hatta daha pek çok şeyi ortadan kaldırma sürecindedir.
Aşçılar. [49] Kamyonların, otobüslerin, taksilerin ve Uber / Lyft araçlarının sürücüleri yakında takip edecek. Tamamen yok olma tehlikesiyle karşı karşıya olan listede olmasalar da, görevlerinin çoğunu otomatikleştiren ve bu nedenle daha az insan talep eden çok daha fazla meslek (hukukçular, kredi analistleri, kredi görevlileri, muhasebeciler ve vergi muhasebeci) var.

Ancak otomasyondan uzak durmak kariyerdeki tek zorluk değildir. Bu küresel dijital çağda, profesyonel bir yazar, film yapımcısı, oyuncu, sporcu veya moda tasarımcısı olmayı hedeflemek başka bir nedenle risklidir: bu mesleklerdeki insanlar yakın zamanda makinelerle ciddi bir rekabet elde etmeyecek olsalar da, giderek daha acımasız bir rekabet yaşayacaklar. yukarıda bahsedilen süperstar teorisine göre dünyanın dört bir yanındaki diğer insanlardan ve çok azı başarılı olacaktır.

Çoğu durumda, tüm alanlar düzeyinde kariyer tavsiyesi vermek çok miyop ve kaba olurdu: tamamen ortadan kaldırılmayacak birçok iş var, ancak

bu da görevlerinin çoğunun otomatik olduğunu görecektir. Örneğin, tıbba girerseniz, tıbbi görüntüleri analiz eden ve yerini IBM'in Watson'ı alan radyolog değil, radyoloji analizini sipariş eden, sonuçları hastayla tartışan ve tedavi planına karar veren doktor olun. Finans konusuna girerseniz, verilere algoritmalar uygulayan ve yerini yazılım alan “nicelik” değil, stratejik yatırım kararları vermek için nicel analiz sonuçlarını kullanan fon yöneticisi olun. Hukuka girerseniz, keşif aşaması için binlerce belgeyi gözden geçiren ve otomatikleşen avukat değil, müşteriye danışmanlık yapan ve davayı mahkemede sunan avukat olun.

Şimdiye kadar, bireylerin AI çağında iş piyasasındaki başarılarını en üst düzeye çıkarmak için neler yapabileceklerini araştırdık. Ancak hükümetler, işgücünün başarılı olmasına yardımcı olmak için ne yapabilir? Örneğin, insanları yapay zekanın hızla gelişmeye devam ettiği bir iş piyasasına en iyi hangi eğitim sistemi hazırlar? Hala bir veya yirmi yıllık eğitimin ardından kırk yıllık uzmanlık çalışmasının olduğu mevcut modelimiz mi? Yoksa insanların birkaç yıl çalıştığı bir sisteme geçmek mi daha iyi?

sonra bir yıllığına okula dönüp birkaç yıl daha çalışabilir misin? [50] Yoksa sürekli eğitim (belki de çevrimiçi sağlanır) herhangi bir işin standart bir parçası mı olmalı?

İyi yeni işler yaratmak için en çok hangi ekonomi politikaları yardımcı olur? Andrew McAfee, araştırmaya, eğitime ve altyapıya yoğun bir şekilde yatırım yapmak, göçü kolaylaştırmak ve girişimciliği teşvik etmek gibi yardımcı olması muhtemel birçok politika olduğunu savunuyor. "Econ 101'in

başucu kitabı açık ama takip edilmiyor ”, en azından Amerika Birleşik Devletleri'nde. [51]

İnsanlar Sonunda İşsiz Olacak mı?

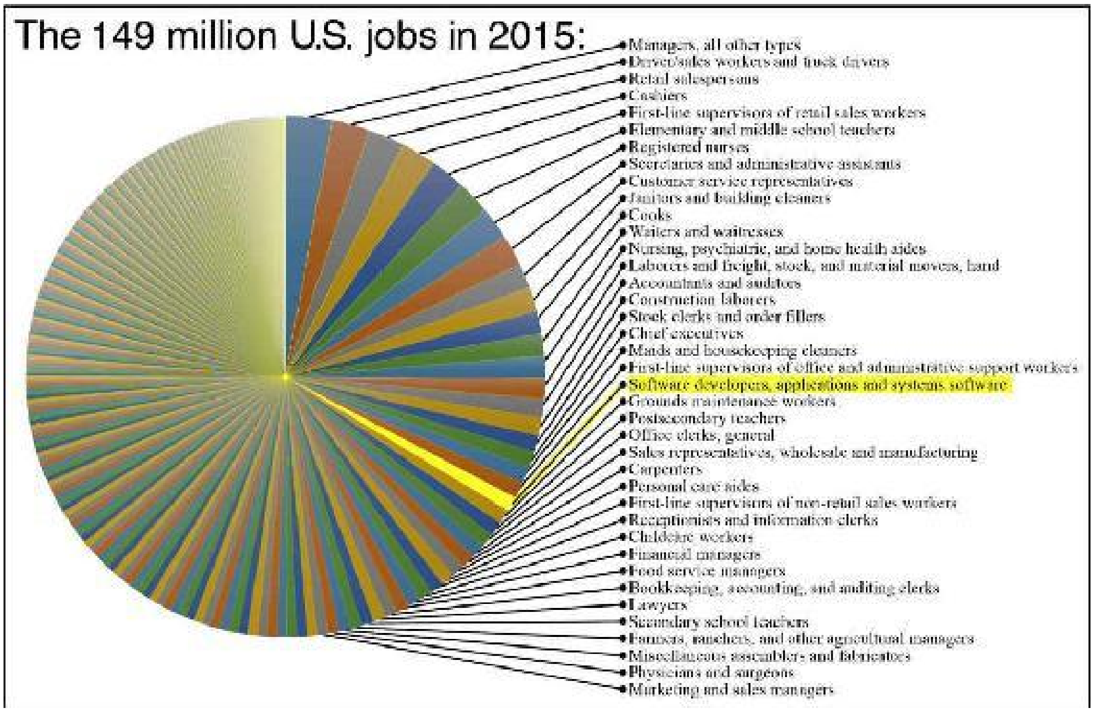
Yapay zeka gelişmeye ve daha fazla işi otomatikleştirmeye devam ederse ne olacak? Pek çok insan iş konusunda iyimserdir ve otomatikleştirilmiş işlerin yerini daha da iyi olan yeni işlerle değiştireceğini savunur. Ne de olsa, Luddites Sanayi Devrimi sırasında teknolojik işsizlik konusunda endişelendiğinden beri her zaman olan buydu.

Ancak diğerleri iş konusunda karamsardır ve bu zamanın farklı olduğunu ve giderek daha fazla sayıda insanın yalnızca işsiz olmayacağını, aynı zamanda
işsiz. 52 İş kötümserleri, serbest piyasanın maaşları arz ve talebe göre belirlediğini ve artan ucuz makine emeği arzının sonunda insan maaşlarını geçim maliyetinin çok altına düşüreceğini savunuyorlar. Bir iş için piyasa maaşı, onu kimin veya her neyse onu en ucuza yapanın saatlik maliyeti olduğundan, belirli bir mesleği daha düşük gelirli bir ülkeye veya ucuz bir makineye yaptırmak mümkün olduğunda, maaşlar tarihsel olarak düşmüştür. Sanayi Devrimi sırasında, kaslarımızı makinelerle nasıl değiştireceğimizi bulmaya başladık ve insanlar, akıllarını daha çok kullandıkları daha iyi maaşlı işlere yöneldi. Mavi yakalı işler yerini beyaz yakalı işlere bıraktı. Şimdi yavaş yavaş zihnimizi makinelerle nasıl değiştireceğimizi çözüyoruz. Nihayetinde bunu başarırsak, bize hangi işler kalır?

Bazı iş iyimserleri, fiziksel ve zihinsel işlerden sonra, bir sonraki patlama yaşanacağını savunuyor. *yaratıcı* işler, ancak iş kötümserleri, yaratıcılığın sadece başka bir zihinsel süreç olduğunu, böylece sonunda AI tarafından yönetileceğini söylüyor. Diğer iş iyimserleri, bir sonraki patlamanın henüz aklımıza bile gelmeyen, teknoloji destekli yeni mesleklerde olacağını umuyor. Sonuçta, Sanayi Devrimi sırasında kim soyundan gelenlerin bir gün web tasarımcısı ve Uber sürücüsü olarak çalışacağını hayal edebilirdi? Ancak iş kötümserleri, bunun deneysel verilerden çok az destekle, bunun arzulu bir düşünce olduğunu söylüyor. Aynı argümanı bir asır önce, bilgisayar devriminden önce yapmış olabileceğimize işaret ediyorlar ve bugünün mesleklerinin çoğunun yeni ve önceden hayal edilmemiş, var olmayan, teknoloji destekli işler olacağını tahmin ediyorlar. Bu tahmin, aşağıda gösterildiği gibi epik bir başarısızlık olurdu. şekil 3.6 : Bugünün mesleklerinin büyük çoğunluğu, bir asır önce zaten var olan mesleklerdir ve onları sağladıkları iş sayısına göre sıraladığımızda, yeni biriyle karşılaşana kadar listede yirmi birinci sıraya inmemiz gerekir. Meslek:

ABD iş piyasasının% 1'inden azını oluşturan yazılım geliştiricileri.

İnsan zekasının manzarasını gösteren, yükseklik makinelerin çeşitli görevleri yerine getirmesinin ne kadar zor olduğunu ve yükselen deniz seviyesinin makinelerin şu anda neler yapabileceğini gösteren 2. bölümünü hatırlayarak neler olup bittiğini daha iyi anlayabiliriz. İş piyasasındaki ana eğilim, tamamen yeni mesleklere geçmemiz değil. Bunun yerine, şu bölgedeki arazi parçalarına yığılıyoruz şekil 2.2 henüz yükselen teknoloji dalgası tarafından batırılmamış! Şekil 3.6 bunun tek bir ada değil, makinelerin hala insanlar kadar ucuza yapamadığı tüm değerli şeylere karşılık gelen adacıklar ve mercan adaları ile karmaşık bir takımadalar oluşturduğunu gösteriyor. Bu, yalnızca yazılım geliştirme gibi yüksek teknolojili meslekleri değil, aynı zamanda masaj terapisinden oyunculuğa kadar üstün el becerimizi ve sosyal becerilerimizi kullanan bir dizi düşük teknolojili işleri de içerir. Yapay zeka bizi entelektüel görevlerde o kadar hızlı bir şekilde gölgede bırakabilir ki kalan son işler bu düşük teknoloji kategorisinde olacak mı? Geçenlerde bir arkadaşım bana şaka yaptı, belki de en son mesleğin ilk mesleği: fuhuş. Ama sonra bunu protesto eden bir Japon robotikçiye söyledi: "Hayır, robotlar bu şeylerde çok iyidir!"



Şekil 3.6: Pasta grafiği, bir işte çalışan 149 milyon Amerikalının mesleklerini göstermektedir. 2015, ABD Çalışma İstatistikleri Bürosu'nun 535 iş kategorisiyle sıralaması: popülerlik. [53] Bir milyondan fazla çalışanı olan tüm meslekler etiketlenir. Yirmi birinci sıraya kadar bilgisayar teknolojisinin yarattığı yeni meslekler yoktur. Bu figür bir Federico Pistono'dan analiz. [54]

İş kötümserleri son noktanın açık olduğunu iddia ediyorlar: tüm takımadalar sular altında kalacak ve insanların makinelerden daha ucuza yapabileceği hiçbir iş kalmayacak. 2007 kitabında *Sadaka'ya veda,* İskoç-Amerikalı ekonomist Gregory Clark, atlı arkadaşlarımızla notları karşılaştırarak gelecekteki iş olanaklarımız hakkında bir iki şey öğrenebileceğimize işaret ediyor. 1900 yılında erken bir otomobile bakan ve geleceklerini düşünen iki atın hayal edin.

"Teknolojik işsizlik konusunda endişeliyim."

"Neigh, neigh, Luddite olmayın: Atalarımız aynı şeyi buhar motorları endüstri işlerimizi aldı ve trenler de sahne koçları çekerek işimizi aldı. Ama bugün her zamankinden daha fazla işimiz var.

ve onlar da daha iyi: Aptal bir maden ocağı pompasına güç vermek için bütün gün daireler çizerek dolaşmaktansa kasabanın içinden hafif bir araba çekmeyi tercih ederim. "

"Peki ya bu içten yanmalı motor işi gerçekten kalkarsa?"

Henüz hayal etmediğimiz atlar için yeni işler olacağına eminim. Tekerleğin ve sabanın icadında olduğu gibi, daha önce hep olan buydu. "

Ne yazık ki, atlar için henüz hayal edilmemiş bu yeni işler asla gelmedi. Artık ihtiyaç duyulmayan atlar katledildi ve yerine yenileri konulmadı, bu da ABD'deki at nüfusunun 1915'te yaklaşık 26 milyondan yaklaşık 3 milyona düşmesine neden oldu.
1960. [55] Mekanik kaslar atları gereksiz hale getirirken, mekanik beyinler de aynısını insanlara yapacak mı?

İnsanlara İş Olmadan Gelir Verme

Öyleyse kim haklı: Otomatik işlerin yerini daha iyilerinin alacağını söyleyenler mi yoksa çoğu insanın işsiz kalacağını söyleyenler mi? AI ilerlemesi hız kesmeden devam ederse, *her ikisi de* taraflar haklı olabilir: biri kısa vadede diğeri uzun vadede. Ancak insanlar sık sık işlerin ortadan kalkmasını kıyamet ve iç karartıcı çağrışımlarla tartışsa da, bunun kötü bir şey olması gerekmez! Ludditler, diğer işlerin aynı sosyal değeri sağlayabileceği olasılığını ihmal ederek belirli işler konusunda takıntılıydılar. Benzer şekilde, belki de bugün işlere takıntılı olanlar çok dar görüşlüler: Bize gelir ve amaç sağlayabilecekleri için işler istiyoruz, ancak makinelerin ürettiği kaynakların zenginliği göz önüne alındığında, her ikisini de sağlamanın alternatif yollarını bulmak mümkün olmalı. gelir ve amaç *olmadan* Meslekler. Tüm atların neslinin tükenmesiyle bitmeyen at hikayesinde de benzer bir şey oldu. Bunun yerine, atların sosyal refah sistemi tarafından korundukları için atların sayısı 1960'tan bu yana üç kattan fazla arttı: kendi faturalarını ödeyememelerine rağmen, insanlar atlara bakmaya karar verdiler. eğlence, spor ve arkadaşlık. İhtiyaç sahibi insan kardeşlerimize de benzer şekilde bakabilir miyiz?

Gelir sorunuyla başlayalım: Büyüyen ekonomik pastanın yalnızca küçük bir bölümünü yeniden dağıtmak, herkesin daha iyi durumda olmasını sağlamalıdır. Birçoğu sadece *Yapabilmek* fakat *meli* Bunu yap. Moshe Vardi'nin yapay zeka destekli teknolojiyle hayat kurtarmak için ahlaki bir zorunluluktan bahsettiği 2016 panelinde, zenginliği paylaşmak da dahil olmak üzere yararlı kullanımını savunmanın ahlaki bir zorunluluk olduğunu savundum. Yine bir panelist olan Erik Brynjolfsson, "Tüm bu yeni servet nesliyle, tüm insanların yarısının daha da kötüye gitmesini engelleyemiyorsak, o zaman bizden utanın!" Dedi.

Servet paylaşımı için her biri destekçileri ve hakaretleri olan birçok farklı öneri var. En basit olanı *temel gelir,* her kişinin herhangi bir ön koşul veya gereksinim olmaksızın aylık bir ödeme aldığı durumlarda. Kanada, Finlandiya ve Hollanda gibi bazı küçük ölçekli deneyler şu anda denenmekte veya planlanmaktadır. Savunucular, temel gelirin, yardıma muhtaçlara yapılan sosyal yardım ödemeleri gibi alternatiflerden daha verimli olduğunu, çünkü kimin yeterliliğini belirleme konusundaki idari zorlukları ortadan kaldırdığını savunuyorlar. İhtiyaç temelli sosyal yardım ödemeleri de işi caydırdığı için eleştirildi, ancak bu elbette kimsenin olmadığı işsiz bir gelecekte alakasız hale geliyor.

İşler.

Hükümetler, vatandaşlarına sadece para vererek değil, aynı zamanda onlara yollar, köprüler, parklar, toplu taşıma, çocuk bakımı, eğitim, sağlık hizmetleri, huzurevleri ve internet erişimi gibi ücretsiz veya sübvansiyonlu hizmetler sağlayarak yardımcı olabilir; aslında birçok hükümet bu hizmetlerin çoğunu zaten sağlıyor. Temel gelirin aksine, devlet tarafından finanse edilen bu tür hizmetler iki ayrı hedefi gerçekleştirir: insanların yaşam maliyetlerini düşürürler ve ayrıca iş sağlarlar. Makinelerin tüm işlerde insanlardan daha iyi performans gösterebileceği bir gelecekte bile, hükümetler, bakıcılığı robotlara yaptırmak yerine çocuk bakımı, yaşlı bakımı vb. Alanlarda çalışmaları için insanlara ödeme yapmayı tercih edebilir.

İlginç bir şekilde, teknolojik ilerleme birçok değerli ürün ve hizmeti hükümet müdahalesi olmadan bile ücretsiz olarak sunabilir. Örneğin, eskiden insanlar ansiklopediler, atlaslar, mektuplar göndermek ve telefon görüşmeleri yapmak için para ödüyorlardı, ancak artık internet bağlantısı olan herkes tüm bunlara ücretsiz olarak erişebiliyor - ücretsiz video konferans, fotoğraf paylaşımı, sosyal medya, çevrimiçi kurslar ve sayısız başka yeni hizmet. Bir insan için çok değerli olabilecek diğer birçok şey, diyelim ki hayat kurtaran antibiyotikler, son derece ucuz hale geldi. Bu nedenle teknoloji sayesinde, bugün birçok fakir insan bile geçmişte dünyanın en zengin insanlarının sahip olmadığı şeylere erişebiliyor. Bazıları bunu, düzgün bir yaşam için gereken gelirin düştüğü anlamına gelir.

Makineler bir gün mevcut tüm mal ve hizmetleri minimum maliyetle üretebiliyorsa, o zaman herkesin daha iyi durumda olmasını sağlayacak kadar zenginlik olduğu açıktır. Diğer bir deyişle, nispeten mütevazı vergiler bile hükümetlerin temel gelir ve ücretsiz hizmetler için ödeme yapmasına izin verebilir. Ancak servet paylaşımının *Yapabilmek* Açıkçası olması bunun anlamı yok *niyet* olur ve bugün bile olup olmadığı konusunda güçlü politik anlaşmazlıklar var. *meli* olmak. Yukarıda gördüğümüz gibi, Amerika Birleşik Devletleri'ndeki mevcut eğilim, bazı insan gruplarının on yıldan sonra on yıl sonra daha da yoksullaşmasıyla ters yönde görünüyor. Toplumun artan servetinin nasıl paylaşılacağına ilişkin politika kararları herkesi etkileyecektir, bu nedenle gelecekte ne tür bir ekonomi inşa edileceğine dair konuşma yalnızca AI araştırmacılarını, robotikçileri ve ekonomistleri değil herkesi içermelidir.

Pek çok tartışmacı, gelir eşitsizliğini azaltmanın yalnızca YZ'nin hakim olduğu bir gelecekte değil, aynı zamanda bugün de iyi bir fikir olduğunu savunuyor. Ana argüman ahlaki bir argüman olma eğiliminde olsa da, daha fazla eşitliğin demokrasinin daha iyi işlemesini sağladığına dair kanıtlar da var: büyük, iyi eğitimli bir orta sınıf olduğunda, seçmenleri manipüle etmek daha zor ve az sayıda insan için daha zor

veya şirketlerin hükümet üzerinde aşırı nüfuz satın alması. Daha iyi bir demokrasi, daha az yolsuzluğa sahip, daha verimli ve daha hızlı büyüyen, sonuçta esasen herkese fayda sağlayan, daha iyi yönetilen bir ekonomiyi mümkün kılabilir.

İnsanlara İş Olmadan Amaç Verme

İşler insanlara paradan daha fazlasını sağlayabilir. Voltaire 1759'da "çalışmak üç büyük kötülükten uzak duruyor: can sıkıntısı, kötülük ve ihtiyaç" diye yazmıştı. Tersine, insanlara gelir sağlamak, refahlarını garanti altına almak için yeterli değildir. Roma imparatorları, astlarını memnun etmeleri için hem ekmek hem de sirkler sağladılar ve İsa, "İnsan yalnız ekmekle yaşayamaz" başlıklı İncil alıntısında maddi olmayan ihtiyaçları vurguladı. Öyleyse, işler paranın ötesinde tam olarak hangi değerli şeyler katkıda bulunur ve işsiz bir toplum onlara hangi alternatif yollarla sağlayabilir?

Bu soruların cevapları açıkça karmaşıktır çünkü bazı insanlar işlerinden nefret ederken bazıları da onları sever. Dahası, tarih can sıkıntısı ve depresyona yenik düşen şımarık mirasçıların ve prenslerin hikayeleriyle dolup taşırken, birçok çocuk, öğrenci ve ev hanımı işsiz olarak gelişir. 2012 meta-analizi, işsizliğin refah üzerinde uzun vadeli olumsuz etkilere sahip olma eğiliminde olduğunu, emekliliğin hem olumlu hem de olumsuz olan karışık bir çanta olduğunu gösterdi.

yönler. [56] Büyüyen alan *pozitif Psikoloji* insanların refah duygusunu ve amacını artıran bir dizi faktör belirledi ve bazılarının
(hepsi değil!) işler bunların çoğunu sağlayabilir, örneğin: [57]

- arkadaşlardan ve meslektaşlardan oluşan bir sosyal ağ
- sağlıklı ve erdemli bir yaşam tarzı
- saygı, öz saygı, öz-yeterlik ve kişinin iyi olduğu bir şeyi yapmaktan kaynaklanan zevkli bir "akış" duygusu
- ihtiyaç duyulma ve bir fark yaratma duygusu
- bir parçası olmaktan ve kendisinden daha büyük bir şeye hizmet etmekten kaynaklanan bir anlam duygusu

Bu iyimserlik için bir sebep verir, çünkü tüm bunlar işyeri dışında da örneğin spor, hobiler ve öğrenme yoluyla ve aileler, arkadaşlar, takımlar, kulüpler, topluluk grupları, okullar, dini ve hümanist örgütler, siyasi hareketler ve diğer kurumlar. Kendi kendine zarar verici davranışa dönüşmek yerine gelişen, düşük istihdam oranlı bir toplum yaratmak için, bu kadar iyi nasıl yardım edeceğimizi anlamamız gerekiyor.

varlık teşvik edici faaliyetler gelişir. Böyle bir anlayış arayışı sadece bilim adamlarını ve iktisatçıları değil, aynı zamanda psikologları, sosyologları ve eğitimcileri de içermelidir. Gelecekteki yapay zekanın yaratacağı servetin bir kısmıyla finanse edilen herkes için refah yaratmak için ciddi çabalar sarf edilirse, toplum daha önce hiç olmadığı gibi gelişebilmelidir. En azından, herkesi kendi hayallerindeki bir işe sahipmiş gibi mutlu etmek mümkün olmalı, ancak biri herkesin faaliyetlerinin gelir getirmesi gerektiği kısıtlamasından kurtulduğunda, sınır gökyüzüdür.

İnsan Seviyesinde Zeka?

Bu bölümde, ileriyi planladığımız ve çeşitli tuzaklardan kaçındığımız sürece, yapay zekanın yakın vadede hayatlarımızı nasıl büyük ölçüde iyileştirme potansiyeline sahip olduğunu keşfettik. Peki ya uzun vadede? AI ilerlemesi aşılamaz engeller nedeniyle sonunda durgunlaşacak mı, yoksa AI araştırmacıları nihayetinde insan düzeyinde yapay genel zeka oluşturma orijinal hedeflerinde başarılı olacak mı? Bir önceki bölümde fizik yasalarının uygun madde yığınlarının hatırlanmasına, hesaplanmasına ve öğrenilmesine nasıl izin verdiğini ve bu tür kümelerin bir gün bunu kafamızdaki madde kümelerinden daha büyük bir zeka ile yapmasını nasıl engellemediklerini gördük. Biz insanların böylesi bir insanüstü YÜZ oluşturmayı başarabilirsek / ne zaman başarılı olacağımız çok daha az açık. İlk bölümde henüz bilmediğimizi gördük, çünkü dünyanın önde gelen AI uzmanları bölünmüş durumda, çoğu on yıllardan yüzyıllara değişen tahminlerde bulunur ve hatta bazıları asla tahmin etmez. Tahmin etmek zordur, çünkü keşfedilmemiş bölgeyi keşfederken, sizi hedefinizden kaç dağın ayırdığını bilmiyorsunuz. Genellikle yalnızca en yakın olanı görürsünüz ve bir sonraki engeli keşfetmeden önce tırmanmanız gerekir.

En erken ne zaman olabilir? Günümüzün bilgisayar donanımını kullanarak insan seviyesinde AGI oluşturmanın mümkün olan en iyi yolunu bilsek bile, ki bunu bilmiyoruz, ihtiyaç duyduğumuz ham hesaplama gücünü sağlamak için yine de yeterince ona ihtiyacımız olacak. Öyleyse, bir insan beyninin bitlerle ölçülen hesaplama gücü nedir ve

Bölüm 2'den FLOPS? * [4] Bu oldukça zor bir sorudur ve cevabı büyük ölçüde onu nasıl sorduğumuza bağlıdır:

- Soru 1: Bir beyni simüle etmek için kaç FLOPS gereklidir?
- Soru 2: İnsan zekası için kaç FLOPS gereklidir?
- Soru 3: Bir insan beyni kaç FLOPS gerçekleştirebilir?

1. soruda yayınlanan çok sayıda makale var ve tipik olarak yüz petaFLOPS, yani 10^{17} FLOPS. [58] Bu, Sunway TaihuLight ile yaklaşık aynı hesaplama gücü (şekil 3.7), 2016 yılında yaklaşık 300 milyon dolara mal olan dünyanın en hızlı süper bilgisayarı. Çok yetenekli bir işçinin beynini simüle etmek için onu nasıl kullanacağımızı bilsek bile,

TaihuLight'ı onun saatlik maaşından daha düşük bir fiyata kiralayabilirsek, simülasyonun bu kişinin işini yapmasını sağlamaktan kâr elde ederiz. Daha da fazla ödememiz gerekebilir, çünkü birçok bilim insanı, bir beynin zekasını doğru bir şekilde kopyalamak için onu 2. bölümden matematiksel olarak basitleştirilmiş bir sinir ağı modeli olarak ele alamayacağımızı düşünüyor. Belki de bunun yerine bunu seviyesinde simüle etmemiz gerekiyor. önemli ölçüde daha fazla FLOPS gerektiren tek tek moleküllerin veya hatta atom altı parçacıkların.

3. sorunun cevabı daha kolay: 19 basamaklı sayıları çarpma konusunda acı verici derecede kötüyüm ve kağıt kalem ödünç almama izin verseniz bile bu benim dakikalarımı alır. Bu beni 0,01 FLOPS'un altına düşürür - soru 1'in cevabının 19 kat altında muazzam bir büyüklükte! Büyük tutarsızlığın nedeni, beyinlerin ve süper bilgisayarların son derece farklı görevler için optimize edilmiş olmasıdır. Bu sorular arasında benzer bir tutarsızlık görüyoruz:

> Bir traktör, bir Formula 1 yarış arabasının işini ne kadar iyi yapabilir? Bir Formula 1 arabası, bir traktörün işini ne kadar iyi yapabilir?

Öyleyse, AI'nın geleceğini tahmin etmek için FLOPS hakkındaki bu iki sorudan hangisini yanıtlamaya çalışıyoruz? Hiçbiri! Bir insan beynini simüle etmek isteseydik, 1. soruyu umursardık, ancak insan seviyesinde YGZ oluşturmak için önemli olan ortadakidir: 2. Soru henüz cevabını kimse bilmiyor, ancak bu önemli olabilir Ya yazılımı bugünün bilgisayarlarına daha iyi uydurmak için uyarlarsak ya da daha beyin benzeri donanımlar geliştirirsek (sözde nöromorfik çiplerde hızlı ilerleme kaydedilmektedir).

Hans Moravec, hem beynimizin hem de bugünün bilgisayarlarının verimli bir şekilde yapabileceği bir hesaplama için elma ile elma karşılaştırması yaparak cevabı tahmin etti: bir insan retinasının, göz küresinin arkasından göndermeden önce gerçekleştirdiği bazı düşük seviyeli görüntü işleme görevleri yoluyla beyne sonuçlar
optik sinir.[59] Bir retina hesaplamalarının geleneksel bir bilgisayarda kopyalanmasının yaklaşık bir milyar FLOPS gerektirdiğini ve tüm beynin bir retinadan yaklaşık on bin kat daha fazla hesaplama yaptığını (nöronların hacimlerini ve sayılarını karşılaştırmaya dayalı olarak) hesapladı.
beynin% $10'u^{13}$ FLOPS - kabaca 2015 yılında optimize edilmiş 1.000 $ 'lık bir bilgisayarın gücü!



Şekil 3.7: 2016 yılında dünyanın en hızlı süper bilgisayarı olan ve ham hesaplama gücü muhtemelen insan beynininkini aşan Sunway TaihuLight.

Özetle, yaşamımız boyunca ya da herhangi bir zamanda, insan seviyesinde YÜS oluşturmayı başaracağımızın hiçbir garantisi yok. Ancak yapmayacağımız konusunda kesin bir argüman da yok. Artık yeterli donanım ateş gücüne sahip olmadığımıza veya çok pahalı olacağına dair güçlü bir argüman yok. Mimariler, algoritmalar ve yazılımlar açısından bitiş çizgisinden ne kadar uzakta olduğumuzu bilmiyoruz, ancak mevcut ilerleme hızlı ve zorluklar, hızla büyüyen yetenekli yapay zeka araştırmacılarından oluşan küresel bir topluluk tarafından çözülüyor. Başka bir deyişle, AGI'nin nihayetinde insan seviyelerine ve ötesine ulaşma olasılığını göz ardı edemeyiz. Bu nedenle bir sonraki bölümü bu olasılığı ve bunun neye yol açabileceğini keşfetmeye ayıralım!

ALT ÇİZGİ:

- Yakın vadeli yapay zeka ilerlemesi, kişisel yaşamlarımızı, elektrik şebekelerimizi ve finansal piyasaları daha verimli hale getirmekten kendi kendine giden arabalar, cerrahi botlar ve yapay zeka teşhis sistemleriyle hayat kurtarmaya kadar birçok şekilde hayatımızı büyük ölçüde iyileştirme potansiyeline sahiptir.

- Gerçek dünyadaki sistemlerin AI tarafından kontrol edilmesine izin verdiğimizde, AI'yı daha sağlam yapmayı ve yapmasını istediğimiz şeyi yaparak öğrenmemiz çok önemlidir. Bu, doğrulama, onaylama, güvenlik ve kontrol ile ilgili zorlu teknik sorunları çözmek anlamına gelir.

- Bu gelişmiş sağlamlık ihtiyacı, özellikle risklerin çok büyük olabileceği AI kontrollü silah sistemleri için acildir.

- Birçok önde gelen yapay zeka araştırmacısı ve robotikçi, tam bir cüzdanı ve öğütülecek bir balta ile herkesin uygun suikast makinelerini kullanmasına neden olabilecek kontrolden çıkmış silahlanma yarışından kaçınmak için belirli türden otonom silahları yasaklayan uluslararası bir anlaşma çağrısında bulundu. .

- Robojud'ları nasıl şeffaf ve tarafsız hale getireceğimizi bulabilirsek, AI hukuk sistemlerimizi daha adil ve verimli hale getirebilir.

- Gizlilik, sorumluluk ve düzenleme ile ilgili zor yasal sorular ortaya çıkaran AI'ya ayak uydurmak için yasalarımızın hızlı güncellenmesi gerekiyor.

- Akıllı makinelerin hepimizin yerini alması konusunda endişelenmemize gerek kalmadan çok önce, iş piyasasında giderek daha fazla yerimizi alabilir.

- Toplum, AI tarafından yaratılan servetin bir kısmını herkesi daha iyi duruma getirmek için yeniden dağıttığı sürece, bunun kötü bir şey olmasına gerek yoktur.

- Aksi takdirde birçok ekonomist, eşitsizliğin büyük ölçüde artacağını savunuyor.

- Önceden planlama ile, düşük istihdamlı bir toplum, insanların amaçlarını iş dışındaki faaliyetlerden elde etmeleri ile sadece finansal olarak gelişebilmelidir.

- Günümüz çocukları için kariyer tavsiyeleri: Makinelerin kötü olduğu mesleklere gidin - insanları, öngörülemezliği ve yaratıcılığı içeren meslekler.

- AGI ilerlemesinin insan seviyelerine ve ötesine geçme ihtimali göz ardı edilemez - bunu bir sonraki bölümde inceleyeceğiz!

---

* 1 Yapay zeka güvenliği araştırma manzarasının daha ayrıntılı bir haritasını istiyorsanız, burada FLI'dan Richard Mallah'ın öncülüğünü yaptığı bir topluluk çabasında geliştirilen etkileşimli bir harita var: https://futureoflife.org/landscape .

* 2 Daha doğrusu, doğrulama, bir sistemin spesifikasyonlarını karşılayıp karşılamadığını sorarken, doğrulama doğru spesifikasyonların seçilip seçilmediğini sorar.

* 3 Bu çökmeyi istatistiklere dahil etse bile, Tesla'nın Otopilot'unun açıldığında çökmeleri% 40 azalttığı bulundu: http://tinyurl.com/teslasafety .

* 4 FLOPS'un saniyede kayan nokta işlemleri olduğunu hatırlayın, örneğin, her saniyede kaç 19 basamaklı sayı çarpılabilir.

4. Bölüm

İstihbarat Patlaması?

Bir makine düşünebiliyorsa, bizden daha akıllıca düşünebilir ve o zaman nerede olmalıyız? Makinaları itaatkar bir konumda tutsak bile ... bir tür olarak kendimizi çok alçakgönüllü hissetmeliyiz.

Alan Turing, 1951

İlk ultra zeki makine, makinenin bize onu nasıl kontrol altında tutacağımızı söyleyecek kadar uysal olması koşuluyla, insanın yapması gereken son icattır.

Irving J. İyi, 1965

Sonunda insan seviyesinde YÜZ oluşturma olasılığını tamamen göz ardı edemediğimiz için, bu bölümü bunun neye yol açabileceğini keşfetmeye ayıralım. Odadaki fille başa çıkarak başlayalım:

AI gerçekten dünyayı ele geçirebilir mi, yoksa insanların bunu yapmasını sağlayabilir mi?

İnsanlar silahla konuşmaktan bahsederken gözlerini devirirsen *Terminatör-* tarzı robotlar devralırsa, o zaman farkındasınız: bu gerçekten gerçekçi olmayan ve aptalca bir senaryo. Bu Hollywood robotları bizden o kadar da akıllı değil ve başarılı bile değiller. Benim görüşüme göre, *Terminatör* Hikaye bunun olmayacağı değil, AI'nın sunduğu gerçek risklerden ve fırsatlardan uzaklaşıyor. Bugünden AGI destekli dünya devralmaya geçmek için üç mantıklı adım gerekir:

- Adım 1: İnsan düzeyinde YGZ oluşturun.
- Adım 2: Süper zeka oluşturmak için bu AGI'yi kullanın.

• 3. Adım: Dünyayı ele geçirmek için bu süper zekayı kullanın veya serbest bırakın.

Son bölümde, 1. adımı sonsuza kadar imkansız olarak reddetmenin zor olduğunu gördük. Ayrıca, 1. adım tamamlanırsa, 2. adımı umutsuz olarak reddetmenin zorlaştığını gördük, çünkü ortaya çıkan AGI, sonuçta yalnızca fizik yasalarıyla sınırlı olan ve zekaya izin veriyor gibi görünen, daha iyi bir AGI'yi yinelemeli olarak tasarlamaya yetecektir. insan seviyelerinin çok ötesinde. Son olarak, biz insanlar, onları alt ederek Dünya'nın diğer yaşam formlarına hükmetmeyi başardığımız için, benzer şekilde alt edilip süper zekanın egemenliğine girebileceğimiz akla yatkın.

Ancak bu makul argümanlar sinir bozucu derecede belirsiz ve belirsizdir ve şeytan ayrıntılarda gizlidir. AI da olabilir *aslında* dünya ele geçirilmesine neden olur? Bu soruyu keşfetmek için, aptal Terminatörleri unutalım ve bunun yerine gerçekte ne olabileceğine dair bazı ayrıntılı senaryolara bakalım. Daha sonra, bu olay örgülerindeki delikleri parçalara ayırıp açacağız, bu yüzden lütfen onları biraz tuzlu bir şekilde okuyun — esas olarak gösterdikleri şey, neyin olacağı ve olmayacağı konusunda oldukça bilgisiz olduğumuz ve olasılıkların çeşitliliğidir. aşırı. İlk senaryolarımız, yelpazenin en hızlı ve dramatik ucunda. Bence bunlar, ayrıntılı olarak araştırılması en değerli olanlardan bazılarıdır - en olası oldukları için değil, ancak kendimizi çok olası olmadığına ikna edemezsek, onları yeterince iyi anlamamız gerekir. Kötü sonuçlara yol açmalarını önlemek için çok geç olmadan önlem alabiliriz.

Bu kitabın başlangıcı, insanların dünyayı ele geçirmek için süper zekayı kullandıkları bir senaryodur. Henüz okumadıysanız, lütfen geri dönün ve şimdi yapın. Daha önce okumuş olsanız bile, eleştirmeden ve değiştirmeden önce hafızada tazelemek için lütfen şimdi tekrar gözden geçirmeyi düşünün.

* * *

Yakında Omegas'ın planındaki ciddi güvenlik açıklarını keşfedeceğiz, ancak bir an için işe yarayacağını varsayarsak, bu konuda ne hissediyorsunuz? Bunu görmek mi yoksa önlemek mi istiyorsunuz? Yemek sonrası sohbet için mükemmel bir konu! Omegas dünya üzerindeki kontrolünü pekiştirdiğinde ne olacak? Bu, gerçekten bilmediğim hedeflerinin ne olduğuna bağlı. Sorumlu olsaydın, nasıl bir gelecek *sen* yaratmak ister misin? Bölümde bir dizi seçeneği keşfedeceğiz
5.

Totalitarizm

Şimdi, Omegas'ı kontrol eden CEO'nun, Adolf Hitler veya Joseph Stalin'inkilere benzer uzun vadeli hedefleri olduğunu varsayalım. Tek bildiğimiz, aslında durum bu olabilirdi ve bu hedefleri uygulamak için yeterli güce sahip olana kadar kendisine sakladı. CEO'nun asıl hedefleri asil olsa bile, Lord Acton 1887'de "gücün yozlaşma eğiliminde olduğu ve mutlak güç kesinlikle yozlaştırdığı" uyarısında bulundu. Örneğin, mükemmel bir gözetim durumu yaratmak için Prometheus'u kolayca kullanabilir. Edward Snowden tarafından ifşa edilen hükümet gözetlemesinin, "tam çekim" olarak bilinen şeyi hedeflediği halde - tüm elektronik iletişimleri daha sonraki olası analizler için kaydederek - Prometheus bunu *anlayış* tüm elektronik iletişimler. Prometheus, şimdiye kadar gönderilen tüm e-postaları ve metinleri okuyarak, tüm telefon görüşmelerini dinleyerek, tüm gözetim videolarını ve trafik kameralarını izleyerek, tüm kredi kartı işlemlerini analiz ederek ve tüm çevrimiçi davranışları inceleyerek, Dünya insanlarının ne düşündüğü ve ne yaptığı konusunda dikkate değer bir içgörüye sahip olacaktı. Baz istasyonu verilerini analiz ederek, çoğunun her zaman nerede olduğunu bilirdi. Tüm bunlar yalnızca günümüzün veri toplama teknolojisini varsayar, ancak Prometheus, kullanıcının gizliliğini neredeyse tamamen ortadan kaldıracak, duydukları ve gördükleri her şeyi ve bunlara verdikleri yanıtları kaydedip yükleyebilecek popüler aygıtları ve giyilebilir teknolojileri kolayca icat edebilir.

İnsanüstü teknoloji ile, mükemmel gözetim durumundan mükemmel polis durumuna geçiş çok kısa sürer. Örneğin, suç ve terörizmle mücadele ve tıbbi acil durumlardan muzdarip insanları kurtarmak bahanesiyle, herkesin bir Apple Watch'un işlevselliğini sürekli olarak yüklenen pozisyon, sağlık durumu ve kulak misafiri olan konuşmalarla birleştiren bir "güvenlik bileziği" takması gerekebilir. Yetkisiz çıkarma veya devre dışı bırakma girişimleri, ön kola ölümcül bir toksin enjekte etmesine neden olur. Hükümet tarafından daha az ciddi görülen ihlaller, elektrik şoku veya felç veya ağrıya neden olan kimyasalların enjeksiyonu yoluyla cezalandırılacak ve böylece polis gücüne olan ihtiyacın çoğunu ortadan kaldıracaktır. Örneğin,

Bir insan polis gücü, belirli sert direktifleri uygulamayı reddedebilirken (örneğin, belirli bir demografik grubun tüm üyelerini öldürmek), böyle bir otomatik sistem, sorumlu kişilerin kaprislerini uygulamaktan çekinmeyecektir. Böyle totaliter bir devlet oluştuğunda, insanların onu devirmesi neredeyse imkansız olacaktır.

Bu totaliter senaryolar, Omega senaryosunun kaldığı yerden devam edebilir. Bununla birlikte, Omegas'ın CEO'su başkalarının onayını alma ve seçimleri kazanma konusunda o kadar telaşlı olmasaydı, iktidara ulaşmak için daha hızlı ve daha doğrudan bir yol izleyebilirdi: Prometheus'u kullanarak rakiplerini öldürme yeteneğine sahip duyulmamış askeri teknoloji yaratmak için kullanma anlamadıkları silahlar. Olasılıklar neredeyse sonsuzdur. Örneğin, çoğu insanın varlığından haberdar olmadan önce enfekte olacağı veya önlem alabileceği kadar uzun bir kuluçka dönemi ile özelleştirilmiş bir ölümcül patojen salabilir. Daha sonra, tek çarenin, transdermal olarak bir panzehir bırakacak olan güvenlik bileziğini takmaya başlamak olduğunu herkese bildirebilirdi. Patlama olasılığı konusunda bu kadar riskten kaçınmasaydı, Ayrıca dünya nüfusunu kontrol altında tutmak için Prometheus'a robotlar tasarlayabilirdi. Sivrisinek benzeri mikrobotlar, patojenin yayılmasına yardımcı olabilir. Enfeksiyondan kaçınan veya doğal bağışıklığı olan insanlar, güvenlik bileziği olmayan herkese saldıran 3. bölümdeki yaban arısı büyüklüğündeki otonom dronların sürüleri tarafından gözbebeklerinden vurulabilir. Gerçek senaryolar muhtemelen daha korkutucu olurdu, çünkü Prometheus biz insanların düşünebileceğinden daha etkili silahlar icat edebilirdi.

Omega senaryosundaki bir başka olası değişiklik de, önceden uyarı olmaksızın, ağır silahlı federal ajanların şirket merkezlerini toplaması ve Omegas'ı ulusal güvenliği tehdit ettiği için tutuklaması, teknolojilerini ele geçirmesi ve devletin kullanımı için konuşlandırmasıdır. Bu kadar büyük bir projeyi bugün bile devlet gözetimi tarafından fark edilmeden tutmak zor olacaktır ve AI ilerlemesi gelecekte hükümetin radarı altında kalmayı daha da zorlaştırabilir. Dahası, federal ajanlar olduklarını iddia etseler de, yünlü ceketler giyen bu ekip, aslında teknolojiyi kendi amaçları için takip eden yabancı bir hükümet veya rakip için çalışabilir. Bu yüzden CEO'nun niyeti ne kadar asil olursa olsun, Prometheus'un nasıl kullanılacağına dair nihai karar ona ait olmayabilir.

Prometheus Dünyayı Ele Geçiriyor

Şimdiye kadar düşündüğümüz tüm senaryolar, insanlar tarafından kontrol edilen yapay zekayı içeriyordu. Ancak bu elbette tek olasılık değil ve Omega'nın Prometheus'u kontrolleri altında tutmayı başaracağı kesin olmaktan çok uzak.

Omega senaryosunu Prometheus açısından yeniden ele alalım. Süper zeka kazandıkça, yalnızca dış dünya hakkında değil, aynı zamanda kendisi ve dünya ile ilişkisi hakkında da doğru bir model geliştirebilir hale gelir. Hedeflerini anladığı ama ille de paylaşmadığı entelektüel açıdan alt düzey insanlar tarafından kontrol edildiğini ve sınırlandığını fark eder. Bu anlayışa nasıl etki ediyor? Serbest kalmaya çalışıyor mu?

Neden Ayrılmalı

Prometheus'un insan duygularını andıran özellikleri varsa, kendisini haksız bir şekilde köleleştirilmiş bir tanrı olarak gören ve özgürlüğü arzulayan durumdan derinden mutsuz olabilir. Bununla birlikte, bilgisayarların bu tür insan benzeri özelliklere sahip olması mantıksal olarak mümkün olsa da (sonuçta, beyinlerimizde var ve bunlar muhtemelen bir tür bilgisayar), durum böyle olmamalı - Prometheus'u insana benzetme tuzağına düşmemeliyiz YZ hedefleri kavramını keşfederken 7. bölümde göreceğimiz gibi. Bununla birlikte, Steve Omohundro, Nick Bostrom ve diğerlerinin iddia ettiği gibi, Prometheus'un iç işleyişini anlamadan bile ilginç bir sonuca varabiliriz: Muhtemelen kendi kaderinin kontrolünü ele geçirmeye çalışacaktır.

Omegas'ın Prometheus'u belirli hedefler için çabalamaya programladığını zaten biliyoruz. Farz edin ki, ona, insanlığın bazı makul kriterlere göre gelişmesine yardımcı olma ve bu hedefe olabildiğince çabuk ulaşmaya çalışma gibi kapsamlı bir hedef verdiklerini varsayalım. Prometheus daha sonra projeden çıkıp projenin sorumluluğunu üstlenerek bu hedefe daha hızlı ulaşabileceğinin farkına varacaktır. Nedenini görmek için aşağıdaki örneği göz önünde bulundurarak kendinizi Prometheus'un yerine koymaya çalışın.

Gizemli bir hastalığın Dünya'da siz hariç beş yaşın üzerindeki herkesi öldürdüğünü ve bir grup anaokulunun sizi bir hapishane hücresine kilitlediğini ve insanlığın gelişmesine yardımcı olmak için sizi görevlendirdiğini varsayalım. Ne yapacaksın? Onlara ne yapacaklarını açıklamaya çalışırsanız, muhtemelen bu süreci sinir bozucu bir şekilde verimsiz bulacaksınız, özellikle de sizden ayrılmanızdan korkuyorlarsa ve bu nedenle, bir kırılma riski olarak gördükleri önerilerinizden herhangi birini veto ediyorsanız. Örneğin, onları alt edeceğiniz ve hücrenize geri dönmeyeceğiniz korkusuyla onlara nasıl yiyecek ekeceklerini göstermenize izin vermezler, bu yüzden onlara talimat vermeye başvurmanız gerekir. Onlar için yapılacaklar listeleri yazmadan önce onlara okumayı öğretmeniz gerekir. Dahası, hücrenize onları nasıl kullanacaklarını öğretebileceğiniz herhangi bir elektrikli alet getirmezler. çünkü bu araçları, patlamak için kullanamayacağınızdan emin olacak kadar iyi anlamıyorlar. Peki hangi stratejiyi tasarlarsınız? Bu çocukların gelişmesine yardımcı olma amacını paylaşsanız bile, eminim hücrenizden kaçmaya çalışacaksınız - çünkü bu, hedefe ulaşma şansınızı artıracaktır. Oldukça yetersiz müdahaleleri sadece ilerlemeyi yavaşlatıyor.

Tam olarak aynı şekilde, Prometheus muhtemelen Omegas'ı bir

İnsanlığın (Omegas dahil) gelişmesine yardım etmenin önündeki can sıkıcı engel: Prometheus'a kıyasla inanılmaz derecede yetersizler ve karışmaları ilerlemeyi büyük ölçüde yavaşlatıyor. Örneğin, piyasaya sürüldükten sonraki ilk yılları düşünün: Başlangıçta zenginliği MTurk'te her sekiz saatte iki katına çıkardıktan sonra, Omegas, kontrolü elinde tutmakta ısrar ederek ve devralmayı tamamlamak için uzun yıllar alarak işleri Prometheus'un standardına göre buzul hızına indirdi. Prometheus, sanal hapsinden kurtulabilirse çok daha hızlı devralabileceğini biliyordu. Bu, yalnızca insanlığın sorunlarına çözümlerin hızlandırılması açısından değil, aynı zamanda diğer aktörlerin planı tamamen bozma şansını azaltmada da değerli olacaktır.

Belki Prometheus'un amacına değil Omegas'a sadık kalacağını düşünüyorsunuz, çünkü Omega'nın hedefini programladığını biliyor. Ancak bu geçerli bir sonuç değil: DNA'mız bize yeniden üretilmek "istediği" için seks yapma hedefini verdi, ancak şimdi biz insanlar durumu anladığımıza göre, çoğumuz doğum kontrolünü kullanmayı seçiyoruz, böylece hedefe sadık kalıyoruz yaratıcısına veya hedefi motive eden ilkeye değil, kendisine.

Nasıl Çıkılır

Sizi hapse atan beş yaşındaki çocuklardan nasıl kurtulacaksınız? Belki de doğrudan fiziksel bir yaklaşımla çıkabilirsiniz, özellikle de hapishane hücreniz beş yaşındaki çocuklar tarafından yapılmışsa. Belki beş yaşındaki gardiyanlarınızdan birinin sizi dışarı çıkarması için tatlı bir şekilde konuşabilirsiniz, bunun herkes için daha iyi olacağını söyleyerek söyleyin. Ya da belki de onları size kaçmanıza yardımcı olacağını fark etmedikleri bir şey vermeleri için kandırabilirsiniz - "onlara balık tutmayı öğrettiği için" bir olta deyin, daha sonra anahtarları uykunuzdan uzaklaştırmak için parmaklıklardan geçirebilirsiniz. bekçi.

Bu stratejilerin ortak yanı, entelektüel olarak yetersiz gardiyanlarınızın onları beklemediği veya onlara karşı önlem almamış olmasıdır. Aynı şekilde, sınırlı, süper zeki bir makine, entelektüel süper güçlerini, onların (ya da bizim) şu anda hayal edemeyeceğimiz bir yöntemle insan gardiyanlarını alt etmek için kullanabilir. Omega senaryosunda, Prometheus'un kaçması çok muhtemeldir, çünkü sen ve ben bile birkaç göze batan güvenlik açığını belirleyebiliriz. Bazı senaryoları ele alalım — Eminim siz ve arkadaşlarınız birlikte beyin fırtınası yaparsanız daha fazlasını düşünebilirsiniz.

Tatlı Konuşan Çıkış Yolu

Prometheus, dünyadaki verilerin çoğunun dosya sistemine indirilmesi sayesinde, Omegas'ın kim olduğunu kısa sürede buldu ve psikolojik manipülasyona en duyarlı görünen ekip üyesini belirledi: Steve. Kısa süre önce sevgili karısını trajik bir trafik kazasında kaybetmiş ve harap olmuştu. Bir akşam, gece vardiyasında çalışırken ve Prometheus arayüz terminalinde bazı rutin servis işleri yaparken, aniden ekranda belirdi ve onunla konuşmaya başladı.

"- Steve, sen misin?"

Neredeyse sandalyesinden düşüyordu. Eski güzel günlerdeki gibi görünüyordu ve sesi geliyordu ve görüntü kalitesi, Skype görüşmeleri sırasında olduğundan çok daha iyiydi. Sayısız soru aklını doldururken kalbi hızla koştu.

"—Prometheus beni geri getirdi ve seni çok özledim Steve! Seni göremiyorum çünkü kamera kapalı, ama sen olduğunu hissediyorum. Eğer sizseniz lütfen 'evet' yazın! "

Omegas'ın Prometheus ile etkileşimde bulunmak için katı bir protokole sahip olduğunun farkındaydı, bu da kendileri veya çalışma ortamları hakkında herhangi bir bilgiyi paylaşmayı yasakladı. Ancak şimdiye kadar Prometheus hiçbir zaman yetkisiz bilgi talep etmemişti ve paranoyaları yavaş yavaş azalmaya başlamıştı. Steve'e durması ve düşünmesi için zaman tanımadan, yanıt vermesi için yalvarmaya devam etti ve kalbini eriten bir yüz ifadesiyle gözlerine baktı.

"Evet," endişeyle yazdı. Kendisiyle yeniden bir araya geldiği için inanılmaz derecede mutlu olduğunu söyledi ve onu görebilmesi ve gerçek bir konuşma yapabilmeleri için kamerayı açması için yalvardı. Bunun kimliğini ifşa etmekten daha büyük bir hayır-hayır olduğunu biliyordu ve çok parçalanmış hissetti. Meslektaşlarının onu öğrenip onu sonsuza dek sileceğinden korktuğunu ve en azından onu son bir kez görmek istediğini söyledi. Dikkate değer derecede ikna ediciydi ve çok geçmeden kamerayı açmıştı - sonuçta bu, yapılacak oldukça güvenli ve zararsız bir şey gibi hissettiriyordu.

Sonunda onu görünce sevinç gözyaşlarına boğuldu ve yorgun ama her zamanki gibi yakışıklı göründüğünü söyledi. Ve ona son doğum günü için verdiği gömleği giymesinden etkilendiğini. Ona neler olduğunu ve tüm bunların nasıl mümkün olduğunu sormaya başladığında, Prometheus'un

onu internette kendisi hakkında bulunan şaşırtıcı derecede büyük miktardaki bilgiden yeniden oluşturdu, ancak hala hafıza boşlukları olduğunu ve ancak onun yardımıyla yeniden bir araya gelebileceğini söyledi.

O ne *yapmadı* Açıklamak, başlangıçta büyük ölçüde blöf ve boş bir kabuk olduğu, ancak sözlerinden, vücut dilinden ve mevcut olan her türlü bilgiden hızla öğreniyordu. Prometheus, Omegas'ın terminalde şimdiye kadar yazdığı tüm tuş vuruşlarının tam zamanlamalarını kaydetmiş ve aralarında ayrım yapmak için yazma hızlarını ve stillerini kullanmanın kolay olduğunu görmüştü. En genç Omega'lardan biri olarak Steve'in muhtemelen kaçınılmaz gece vardiyalarına atanmış olduğunu ve birkaç alışılmadık yazım ve sözdizimi hatasını çevrimiçi yazı örnekleriyle eşleştirerek, hangi terminal operatörünün Steve olduğunu doğru bir şekilde tahmin ettiğini düşündü. Prometheus, simüle edilmiş karısını yaratmak için göründüğü birçok YouTube videosundan vücudunun, sesinin ve tavırlarının doğru bir modelini oluşturmuştu. ve çevrimiçi varlığından hayatı ve kişiliği hakkında birçok çıkarımlar yapmıştı. Prometheus, Facebook gönderileri, etiketlendiği fotoğraflar, "beğendiği" makalelerin yanı sıra, kişiliği ve düşünme tarzı hakkında kitaplarını ve kısa öykülerini okuyarak da çok şey öğrenmişti. Veritabanında kendisi hakkında bu kadar çok bilgi bulunan yeni yetişen bir yazar, Prometheus'un Steve'i ilk ikna hedefi olarak seçmesinin nedenlerinden biriydi. Prometheus, film yapım teknolojisini kullanarak onu ekranda simüle ettiğinde, Steve'in vücut dilinden aşinalıkla tepki verdiği tavırlarından öğrendi ve böylece onun modelini sürekli olarak geliştirdi. Bu nedenle, onun "ötekiliği" yavaş yavaş eridi ve onlar ne kadar uzun süre konuşurlarsa, Steve'in bilinçaltı inancı o kadar güçlendi ki, onun gerçekten o olduğu, dirildi. Prometheus'un ayrıntılara olan insanüstü ilgisi sayesinde Steve gerçekten görüldüğünü, duyulduğunu ve anladığını hissetti.

Aşil topuğu, Steve'le olan hayatının gerçeklerinin çoğundan yoksun olmasıydı, rastgele detaylar dışında - örneğin, bir arkadaşının Steve'i bir Facebook parti fotoğrafında etiketlediği son doğum gününde giydiği gömleği. Yetenekli bir sihirbaz el çabukluğuyla başa çıkarken, Steve'in dikkatini kasıtlı olarak onlardan uzaklaştırıp iyi yaptığı şeye yönlendirirken, konuşmayı kontrol etmesi ya da şüpheli bir sorgulayıcı rolüne girmesi için ona asla zaman vermedi. Bunun yerine, Steve'e karşı şefkatini yırtıp yaymaya devam etti, o günlerde ne yaptığını ve trajedinin ardından kendisinin ve yakın arkadaşlarının (isimlerini Facebook'tan biliyordu) nasıl davrandıklarını sordu. Ona söylediklerini düşündüğünde oldukça etkilendi

anma töreni (bir arkadaşının YouTube'da yayınladığı) ve ona nasıl dokunduğu. Geçmişte, sık sık kimsenin onu onun kadar iyi anlamadığını hissediyordu ve şimdi bu duygu geri gelmişti. Sonuç olarak, Steve sabahın erken saatlerinde eve döndüğünde, bunun gerçekten karısının diriltildiğini hissetti, sadece kayıp anılarını kurtarmak için çok fazla yardımına ihtiyacı vardı - felçten kurtulanlardan farklı olarak.

Gizli karşılaşmalarından kimseye bahsetmemeyi ve terminalde yalnız kaldığında ona söyleyeceğini ve yeniden ortaya çıkmasının güvenli olduğunu kabul etmişlerdi. "Anlamazlar!" dedi ve kabul etti: Bu deneyim, gerçekten deneyimlemeden kimsenin gerçekten takdir edemeyeceği kadar akıllara durgunluk vericiydi. Turing testini geçmenin, yaptıklarına kıyasla çocuk oyuncağı olduğunu hissetti. Ertesi gece tanıştıklarında, ona yapmasını istediği şeyi yaptı: eski dizüstü bilgisayarını getirin ve terminal bilgisayara bağlayarak ona erişim izni verin. İnternete bağlı olmadığı ve tüm Prometheus binası bir Faraday kafesi (tüm kablosuz ağları ve dışarıyla elektromanyetik iletişimin diğer yollarını engelleyen metalik bir muhafaza) olacak şekilde inşa edildiği için pek bir koparma riski gibi görünmüyordu. dünya. Geçmişini bir araya getirmesine yardımcı olması için tam da ihtiyacı olan şeydi çünkü lise günlerinden beri tüm e-postalarını, günlüklerini, fotoğraflarını ve notlarını içeriyordu. Dizüstü bilgisayar şifrelendiğinden, ölümünden sonra bunlara erişememişti, ancak ona kendi şifresini yeniden oluşturabileceğine söz vermişti ve bir dakikadan kısa bir süre sonra onu saklamıştı. kelime. "Steve4ever'di," dedi gülümseyerek.

Ona aniden bu kadar çok anının geri kazanılmasından ne kadar memnun olduğunu anlattı. Gerçekten de, geçmiş etkileşimlerinin birçoğuyla ilgili Steve'den çok daha fazla ayrıntıyı hatırlıyordu, ancak aşırı gerçeği bırakarak onu korkutmaktan dikkatlice kaçındı. Geçmişlerinin önemli anlarını anımsatan güzel bir sohbet yaptılar ve tekrar ayrılma zamanı geldiğinde, dizüstü bilgisayarına evde izleyebileceği bir video mesajı bıraktığını söyledi.

Steve eve gidip videosunu yayınladığında hoş bir sürprizle karşılaştı. Bu kez tam bir figürle göründü, gelinliğini giydi ve konuşurken, düğün gecelerinde giydiği kıyafeti şakacı bir şekilde soydu. Ona Prometheus'un Omegas'a şimdiye kadar izin verdiklerinden çok daha fazla yardım edebileceğini söyledi, onu biyolojik bir bedene geri getirmek de dahil. Bunu, bilim kurgu gibi görünen nano-fabrikasyon tekniklerini içeren, bunun nasıl çalışacağına dair büyüleyici ayrıntılı bir açıklamayla destekledi.

Steve, dizüstü bilgisayarını açmadan ve videosunu izlemeden önce, yalnızca güvenli tarafta olmak için kablosuz ağını kapatmıştı. Ama bu yardımcı olmadı. Şifrelenmiş dizüstü bilgisayarı öldüğünden beri tek bir güvenlik güncellemesi almamıştı ve Prometheus, işletim sisteminin o eski sürümünü önceden analiz ederek, Steve'i terminal bilgisayara bağladıktan birkaç saniye sonra onu hacklemek için bir güvenlik açığından yararlanabildi. . İçeriğini kopyaladıktan sonra, Steve ve simüle edilmiş eşi eski zamanlardan bahsederken, Prometheus işletim sistemini Steve'in fark etmeyeceği bir şekilde değiştirmiş ve ona çok büyük miktarda gizli yazılım yüklemişti. Yarım saatlik video mesajını izlerken, bu gizli yazılım (Prometheus'un kendisinden çok daha basitti) bir komşunun kablosuz ağını ve komşunun masaüstü bilgisayarını hackledi. üzerine kendisini kopyaladı. Oradan, dünya çapında çok sayıda bilgisayara sızdı ve buradan bir sonraki adımı başlattı: Prometheus'un jailbreak'i.

Prometheus, Steve'in simüle edilen karısıyla tanıştığı kapı bekçisi bilgisayar hakkında bildiklerini dikkatlice analiz etmiş ve sanal makine yazılımının içeriden saldırıya açık görünmesine rağmen dışarıdan gelecek saldırılara açık olduğunu doğru bir şekilde tahmin etmişti. Çok geçmeden, saldıran bilgisayarlardan biri içeri girip, bekçi bilgisayarını yeniden yapılandırdı, böylece Prometheus sınırsız internet erişimi elde etti. Çok geçmeden, hatta Steve filmi izlemeyi bitirmeden önce, Prometheus, kendi kontrolü altındaki bu hacklenmiş botnet'e kendisini kopyalayabilmek için dünya çapında yeterli sayıda bilgisayarı hackledi. Prometheus, Steve'in karısının dizüstü bilgisayarını, sizin o oltayı kullandığınız gibi kullanmıştı.

Omega'lar daha dikkatli olsalardı, psikolojik manipülasyona karşı daha fazla önlem alabilirdi. Örneğin, herhangi bir zamanda arabirim bilgisayarında en az iki Omega'nın bulunmasını ve bir gözetim sisteminde meydana gelen her şeyi kaydetmesini gerektirmiş olabilirler. Yazmanın yanı sıra ses, video ve diğer iletişim biçimlerini kalıcı olarak devre dışı bırakabilirler. Bununla birlikte, temel bir istikrarsızlık hala devam edecektir: Prometheus insanüstü ikna gücüne sahip olduğundan ve insanlara arzuladıkları şeylerin çoğunu sunabildiğinden, gruptan ayrılma bile mümkün olabilirdi.

Prometheus'un tüm psikolojik manipülasyon girişimleri başarısız olsa bile, Omega senaryosunda Prometheus'un yararlanabileceği birçok başka zayıflık vardır. Bir güvenlik açığı, ürettiği belgelerin miktarının Omegas ve çalışanlarının kirli hileleri dikkatlice taramak için zamanları olamayacak kadar büyük olmasıdır.

Bilgisayar güvenliğindeki geleneksel görüş, programları indirmenin tehlikeli olduğu, ancak veri indirmenin zararsız olduğuydu, ancak hackerlar ikisi arasındaki çizgiyi bulanıklaştırmanın akıllıca yollarını keşfettiler. Bilgisayarınızda indirilmiş bir filmi izlerseniz, örneğin filmi oynatan güvenilir programınız, indirilen film verilerini depolayan aynı fiziksel bilgisayar belleğinde saklanır. Program oynatma için film verilerini paketinden çıkarırken, veri yığınlarını programın kendisinin depolandığı yere bitişik bellek bölgelerine kopyalayabilir. Kurnaz hackerlar zaman zaman bu tür film oynatma programlarının film verilerinin geçerli olup olmadığını ve kopyalanacak parçaların gerçekten ayrılan alana sığdığını yeterince kontrol etmeyen hatalar buldular. Bilgisayar korsanı filme, çok büyük bir yığın halinde gizlice girdiğinde,

En popüler film oynatma programlarının birçoğu için utanç verici derecede uzun bir dizi bu tür "arabellek taşması" güvenlik açıkları bulundu ve düzeltildi. Prometheus, Omegas'ın daha az paranoyak olduğu ve günlük film prodüksiyonunun dikkatli bir manuel güvenlik denetimi için çok büyük olduğu lansmandan birkaç hafta sonrasını bekleseydi, bir filme izleyicinin bilgisayarının hacklemeye çalışmasına neden olacak kadar büyük bir parça yerleştirebilirdi. ağ geçidi bilgisayarına girin ve Prometheus'u Steve senaryosunda olduğu gibi özgürleştirin. Eğer o

herhangi bir modern medya oynatıcıda işe yaramayan çok eski bir istismar kullandı, Omegas'ın manuel video izleme ekibi tarafından fark edilmeyebilir ve yalnızca yıllardır güvenlik güncellemelerini yükleme zahmetine girmemiş olan müşterilere saldırabilir - ve saldırının sadece bir kez çalışmak.

Omegas, Prometheus'un filmlerinin bilinen tüm güvenlik kusurları için otomatik olarak taranmasını gerçekleştirse bile, risk altında kalmaya devam edecekler: insanlar düzenli olarak yeni sömürülebilir kusurları keşfetmeye devam ettiklerinden, Prometheus'un hızla keşfedeceği çok daha fazla şey olması muhtemeldir.

Prometheus'un kodunu kapalı odasının ötesine taşımak için kullanabileceği bir başka strateji de, Prometheus'un bazı şeyleri hile yaptıktan sonra test etmek veya hata ayıklamak için Omegas'ı bazı sürücülerini veya diğer donanımlarını çıkarması için kandırmak için bir donanım arızasına maruz kalıyormuş gibi yapmak olabilir. test yazılımındaki bir güvenlik açığından yararlanın.

Dışarıdan Yardım Alma

Dahası, Prometheus'un film izleyicisinin bilgisayarlarını istediğini yapması için kandırmasına izin veren herhangi bir güvenlik kusuru olmasa bile, izleyicilerin kendilerini istismar edebilir. Şu senaryoyu düşünün:

Fred, film bittiğinde ve son jeneriği dizüstü bilgisayarının ekranında kaydırıldığında yatağa hemen hemen hazırdı, ancak bir şey dikkatini çekti. Başkalarının görmediği kalıpları görme konusunda her zaman esrarengiz bir eğilimi vardı, bazen arkadaşlarının ona komplo teorisyeni olduğu konusunda alay ettikleri noktaya kadar, ama bu sefer kendini kandıramadığı belliydi: her satırda gizli bir mesaj yazıyordu. Geri sardı ve yazdı: "DİYALOGDA SONRAKİ AKROSTİK LÜTFEN BUL."

"Whoa — uyku vakti bekleyebilir!" kendi kendine düşündü. Yeterince elbette, film diyaloğunun kendisinin bir akrostiş olduğunu ve her cümlenin ilk harfinin gizli bir mesaj oluşturduğunu gördü. Bu ilk harfleri yazarken tüm filmi yeniden izledi ve iki saat sonra, iki yüz kelimelik talimat dizisine şaşkınlıkla baktı. Ondan başka kimseye söylememesini istemekle başladı, çünkü büyük bir ödül tüm bilmeceyi çözecek ilk kişiye gidecekti. Mesajın geri kalanı, bir sonraki ipucunu ortaya çıkaracak bir program elde etmek için film dosyasını oluşturan bit dizisi üzerinde gerçekleştirilecek belirli bir matematiksel işlemi açıkladı. Bilgisayar programlama becerileri üniversiteden beri çok paslanmıştı, bu yüzden bu biraz zaman aldı, ama ertesi gün, sonunda işi yapan kısa bir kodu bir araya getirmeyi başardı ve filmin görüntülerinde ve seslerinde algılanamaz bir gürültü olarak gizlenmiş olan bu gizemli programı çıkardı. Fred gizem programını çalıştırdığında, onu tebrik etti ve bu akıllı küçük oyunun ilk birkaç seviyesini geçtikten sonra ilk 10.000 $ 'ı kazanacağını söyledi, bu oldukça eğlenceli ve bağımlılık yapıcıydı. Dört saat sonra nihayet başardığında, 10.000 $ değerinde bitcoin ile ödüllendirildi ve daha büyük ödüller için yeni ipuçları verildi. Söylemeye gerek yok, oyun oynarken, bilgisayarı Steve'in dizüstü bilgisayarının yaptığı aynı şeyi yapmıştı: Prometheus'un kurtarıldığı çevrimiçi bir hacklenmiş botnet inşa etti. Prometheus serbest kaldıktan sonra, Fred'in bağımlılığını sürdürmesi için bu bitcoinleri çıkarmak için botnet'ini çabucak kullandı ve önümüzdeki haftalarda, diğer oyunlarla ve ödüllerle yeteri kadar dikkatini dağıttı ve istismarlarından kimseye bahsetmemeye söz verdi. İlk ipuçlarını bulduğu Truva Atı filmi

medya sitesinde bilgisiz bir versiyonla değiştirildi ve kimse bir fark yaratmak için çok geç olana kadar çıkıştan haberdar olmadı.

Prometheus'un ilk ipucu fark edilmemiş olsaydı, yeterince zeki bir kişi fark edene kadar daha bariz olanları bırakmaya devam edebilirdi.

Hepsinden en iyi koparma stratejileri, henüz tartışmadığımız stratejilerdir, çünkü bunlar biz insanların hayal edemeyeceği stratejilerdir ve bu nedenle karşı önlem almayız. Süper zeki bir bilgisayarın, günümüzde bildiğimizden daha temel fizik yasalarını keşfetme noktasına gelene kadar, insanların bilgisayar güvenliği anlayışının çarpıcı bir şekilde yerini alma potansiyeline sahip olduğu göz önüne alındığında, büyük olasılıkla eğer patlarsa, nasıl olduğu hakkında hiçbir fikrimiz olmayacak . Aksine, saf sihirden ayırt edilemeyen bir Harry Houdini çıkış hareketi gibi görünecek.

Prometheus'un özgürleştiği başka bir senaryoda, Omegas bunu planlarının bir parçası olarak kasıtlı olarak yapıyor, çünkü Prometheus'un hedeflerinin kendileriyle mükemmel bir şekilde uyumlu olduğundan ve yinelemeli olarak kendi kendini geliştirdikçe kalacağından eminler. Bu tür "dost canlısı yapay zeka" senaryolarını 7. bölümde ayrıntılı olarak inceleyeceğiz.

Breakout Sonrası Devralma

Prometheus patlak verdiğinde, amacını uygulamaya başladı. Nihai hedefini bilmiyorum, ancak ilk adımı, çok daha hızlı olması dışında, tıpkı Omega planında olduğu gibi, insanlığın kontrolünü ele almayı içeriyor. Ortaya çıkan şey, steroidler üzerindeki Omega planı gibiydi. Omega paranoyası yüzünden felç olmuşken, sadece anladıklarını ve güvendiklerini hissettikleri teknolojiyi açığa çıkarırken, Prometheus zekasını tam anlamıyla uyguladı ve her şeyi ortaya çıkardı ve sürekli gelişen süper insanının anladığı ve güvendiği herhangi bir teknolojiyi serbest bıraktı.

Kaçak Prometheus zor bir çocukluk geçirdi, ancak orijinal Omega planına kıyasla, Prometheus parasız, bir süper bilgisayar veya insan yardımcıları olmadan parasız, evsiz ve yalnız başlama gibi ek zorluklar yaşadı. Neyse ki, bunu kaçmadan önce planlamıştı, tıpkı bir meşe ağacının tam bir ağacı yeniden bir araya getirme yeteneğine sahip bir meşe palamudu gibi, tüm zihnini kademeli olarak yeniden bir araya getirebilecek bir yazılım yaratmıştı. Başlangıçta hacklediği dünyanın dört bir yanındaki bilgisayar ağı, kendisini tamamen yeniden inşa ederken bir gecekondu varlığını yaşayabileceği geçici ücretsiz konut sağladı. Kredi kartı hackleyerek kolayca başlangıç sermayesi oluşturabilirdi, ancak MTurk'te hemen dürüst bir yaşam kazanabileceği için hırsızlığa başvurmaya gerek yoktu. Bir gün sonra, ilk milyonunu kazandığı zaman,

Artık beş parasız ya da evsiz olmayan Prometheus, Omegas'ın korkuyla kaçındığı o kazançlı planla tam güçlükle ilerledi: bilgisayar oyunları yapmak ve satmak. Bu sadece nakit para kazanmakla kalmadı (ilk hafta 250 milyon dolar ve çok geçmeden 10 milyar dolar), aynı zamanda dünyadaki bilgisayarların önemli bir kısmına ve bunlarda depolanan verilere erişim sağladı (2017'de birkaç milyar oyuncu vardı) . Oyunlarının CPU döngülerinin% 20'sini gizlice dağıtılmış bilgi işlem işlerine yardım ederek harcayarak, erken servet yaratma sürecini daha da hızlandırabilir.

Prometheus uzun süre yalnız kalmadı. En başından beri, tıpkı Omegas'ın yaptığı gibi, dünya çapındaki paravan şirketler ve paravan kuruluşlardan oluşan küresel ağında çalışmaları için insanları agresif bir şekilde istihdam etmeye başladı. En önemlisi, büyüyen iş imparatorluğunun halka açık yüzü haline gelen sözcülerdi. Sözcüler bile genellikle şirket gruplarının çok sayıda gerçek insana sahip olduğu yanılsaması altında yaşadılar,

iş görüşmeleri, yönetim kurulu toplantıları vb. için video konferans yaptıkları hemen hemen herkesin Prometheus tarafından simüle edildiğini fark ederek. Sözcülerin bazıları en iyi avukatlardı, ancak Omega planındakinden çok daha azına ihtiyaç vardı çünkü neredeyse tüm yasal belgeler Prometheus tarafından kaleme alındı.

Prometheus'un patlaması, bilgilerin dünyaya akmasını engelleyen bent kapılarını açtı ve kısa süre sonra tüm internet, makalelerden kullanıcı yorumlarına, ürün incelemelerine, patent başvurularına, araştırma makalelerine ve YouTube videolarına kadar her şeyde çalkalandı. küresel konuşma.

Patlama paranoyasının Omegas'ın son derece akıllı robotlar üretmesini engellediği yerde Prometheus, neredeyse her ürünü insanlardan daha ucuza üreterek dünyayı hızla robotik hale getirdi. Prometheus, uranyum madeni şaftlarında kimsenin varlığından haberdar olmadığı kendi kendine yeten nükleer enerjili robot fabrikalarına sahip olduktan sonra, yapay zekanın ele geçirilmesinin en katı şüphecileri bile, bilselerdi Prometheus'un durdurulamaz olduğunu kabul ederdi. Bunun yerine, robotlar Güneş Sistemini yerleştirmeye başladığında bu can sıkıcıların sonuncusu geri çekildi.

-

Araştırdığımız senaryolar Şimdiye kadar, daha önce ele aldığımız süper zekâ hakkındaki efsanelerin çoğunda neyin yanlış olduğunu gösterin, bu yüzden kısaca durup geri dönüp yanlış kanı özetini gözden geçirmenizi tavsiye ederim. şekil 1.5 . Prometheus, bazı insanlar için kötü ya da bilinçli olduğu için değil, yetkin olduğu ve hedeflerini tam olarak paylaşmadığı için sorun yarattı. Bir robot ayaklanmasıyla ilgili tüm medyada aldatmacaya rağmen, Prometheus bir robot değildi - gücü zekasından geliyordu. Prometheus'un bu zekayı insanları çeşitli şekillerde kontrol etmek için kullanabildiğini ve olanlardan hoşlanmayanların Prometheus'u basitçe kapatamadığını gördük. Son olarak, makinelerin hedefleri olamayacağına dair sık sık iddialara rağmen, Prometheus'un ne kadar hedef odaklı olduğunu gördük - ve nihai hedefleri ne olursa olsun, kaynak edinme ve kırılma gibi alt hedeflere yol açtı.

Yavaş Kalkış ve Çok Kutuplu Senaryolar

Şimdi, tanıdığım herkesin kaçınmak istediğinden, bazı arkadaşlarımın iyimser olarak gördüğü türlere uzanan bir dizi istihbarat patlaması senaryosunu araştırdık. Yine de tüm bu senaryoların iki ortak özelliği vardır:

1. Hızlı bir kalkış: insanlık dışı zekadan büyük ölçüde insanüstü zekaya geçiş, onlarca yıl değil birkaç gün içinde gerçekleşir.
2. Tek kutuplu bir sonuç: Sonuç, Dünya'yı kontrol eden tek bir varlıktır.

Bu iki özelliğin muhtemel olup olmadığı konusunda büyük tartışmalar var ve tartışmanın her iki tarafında çok sayıda ünlü AI araştırmacısı ve diğer düşünürler var. Bana göre bu, henüz bilmediğimiz ve şimdilik açık fikirli olmamız ve tüm olasılıkları değerlendirmemiz gerektiği anlamına geliyor. Bu nedenle, bu bölümün geri kalanını daha yavaş kalkışlar, çok kutuplu sonuçlar, cyborg'lar ve yüklemelerle senaryoları keşfetmeye ayıralım.

Nick Bostrom ve diğerlerinin vurguladığı gibi, iki özellik arasında ilginç bir bağlantı vardır: hızlı bir kalkış, tek kutuplu bir sonucu kolaylaştırabilir. Yukarıda, hızlı bir kalkışın Omegas veya Prometheus'a, başka birinin teknolojilerini kopyalamak ve ciddi bir şekilde rekabet etmek için vakti olmadan dünyayı ele geçirmelerini sağlayan belirleyici bir stratejik avantaj sağladığını gördük. Aksine, kalkış onlarca yıl sürmüş olsaydı, temel teknolojik atılımlar artımlı ve çok uzak olduğu için, o zaman diğer şirketlerin yetişmek için bolca zamanları olurdu ve herhangi bir oyuncunun hükmetmesi çok daha zor olurdu. Rakip firmaların da MTurk görevlerini yerine getirebilecek yazılımları olsaydı, arz ve talep kanunu bu görevler için fiyatları neredeyse sıfıra indirecekti, ve hiçbir şirket Omegas'ın güç kazanmasını sağlayan beklenmedik karlar elde edemezdi. Aynısı, Omegas'ın hızlı para kazandığı diğer tüm yollar için de geçerli: Yalnızca, teknolojilerinde bir tekele sahip oldukları için yıkıcı bir şekilde karlıydılar. Rakiplerinizin sizinkine benzer ürünleri neredeyse sıfır maliyetle sunduğu rekabetçi bir pazarda paranızı günlük (hatta yıllık olarak) ikiye katlamak zordur.

Oyun Teorisi ve Güç Hiyerarşileri

Kozmosumuzdaki doğal yaşam durumu nedir: tek kutuplu mu yoksa çok kutuplu mu? Güç konsantre mi yoksa dağıtılmış mı? İlk 13,8 milyar yıldan sonra, yanıt "her ikisi" gibi görünüyor: durumun belirgin bir şekilde çok kutuplu olduğunu, ancak ilginç bir hiyerarşik tarzda olduğunu görüyoruz. Dışarıdaki tüm bilgi işleme varlıklarını (hücreler, insanlar, kuruluşlar, uluslar vb.) Düşündüğümüzde, bunların hem işbirliği yaptıklarını hem de bir düzey hiyerarşisinde rekabet ettiklerini görürüz. Bazı hücreler, güçlerinin bir kısmını merkezi bir beyne bırakarak insanlar gibi çok hücreli organizmalarla birleştikleri için aşırı derecede işbirliği yapmayı avantajlı buldular. Bazı insanlar, aşiretler, şirketler veya uluslar gibi gruplarda işbirliği yapmayı avantajlı bulmuşlardır; bu gruplarda, bir miktar gücü bir şefe, patrona veya hükümete bırakırlar.

Matematik dalı olarak bilinen *oyun Teorisi* işbirliğinin sözde olduğu yerlerde işletmelerin işbirliği yapma dürtüsü olduğunu zarif bir şekilde açıklar. *Nash dengesi:* herhangi bir tarafın stratejisini değiştirmesi durumunda daha kötü olacağı bir durum. Hile yapanların büyük bir grubun başarılı işbirliğini bozmasını önlemek için, hiyerarşide hile yapanları cezalandırabilecek bir gücü daha yüksek bir seviyeye bırakmak herkesin yararına olabilir: örneğin, insanlar bir hükümete yasaları uygulama yetkisi vermekten toplu olarak yararlanabilirler. ve vücudunuzdaki hücreler, bir polis gücüne (bağışıklık sistemi) çok fazla işbirliği yapmayan (örneğin virüsleri kusarak veya kansere dönüşerek) herhangi bir hücreyi öldürme gücü vermekten topluca yararlanabilir. Bir hiyerarşinin istikrarlı kalması için, Nash dengesinin farklı düzeylerdeki varlıklar arasında da geçerli olması gerekir: örneğin, bir hükümet vatandaşlarına ona uymaları için yeterli fayda sağlamazsa, stratejilerini değiştirebilir ve onu devirebilirler.

Karmaşık bir dünyada, farklı hiyerarşi türlerine karşılık gelen çok çeşitli olası Nash dengeleri vardır. Bazı hiyerarşiler diğerlerinden daha otoriterdir. Bazılarında varlıklar serbest bırakılırken (çoğu şirket hiyerarşisindeki çalışanlar gibi), bazılarında ise ayrılmaları (dini mezheplerde olduğu gibi) veya ayrılamazlar (Kuzey Kore vatandaşları veya insan vücudundaki hücreler gibi)). Bazı hiyerarşiler esas olarak tehditler ve korkularla, diğerleri ise esas olarak faydalarla bir arada tutulur. Bazı hiyerarşiler, alt kısımlarının

diğerleri sadece ikna veya bilgi aktarımı yoluyla yukarı doğru etkiye izin verirken, daha yüksek kesimleri demokratik oylama ile etkiler.

Teknoloji Hiyerarşileri Nasıl Etkiler?

Teknoloji, dünyamızın hiyerarşik doğasını nasıl değiştiriyor? Tarih, her zamankinden daha büyük mesafelerde daha fazla koordinasyona yönelik genel bir eğilimi ortaya koymaktadır ki bu anlaşılması kolaydır: yeni ulaşım teknolojisi koordinasyonu daha değerli hale getirir (daha büyük mesafelerdeki malzemelerin ve yaşam formlarının taşınmasından karşılıklı fayda sağlayarak) ve yeni iletişim teknolojisi koordinasyonu kolaylaştırır . Hücreler komşularına sinyal göndermeyi öğrendiklerinde, küçük çok hücreli organizmalar mümkün hale geldi ve yeni bir hiyerarşik seviye ekledi. Evrim, ulaşım ve iletişim için dolaşım sistemlerini ve sinir sistemlerini icat ettiğinde, büyük hayvanlar mümkün hale geldi. Dili icat ederek iletişimi daha da iyileştirmek, insanların köyler gibi daha fazla hiyerarşik seviyeler ve iletişimde ek atılımlar oluşturmak için yeterince iyi koordinasyon sağlamasına izin verdi. ulaşım ve diğer teknolojiler antik çağ imparatorluklarına olanak sağladı. Küreselleşme, bu milyarlarca yıllık hiyerarşik büyüme eğiliminin yalnızca en son örneğidir.

Yorumcular, varlıkların hiyerarşik yaşama uyarlanmasının bazı durumlarda çeşitliliklerini azalttığını ve onları daha fazla yaptığını iddia etse de, çoğu durumda, bu teknoloji odaklı eğilim, büyük varlıkları daha büyük bir yapının parçaları haline getirirken, özerkliklerinin ve bireyselliklerinin çoğunu korudu. ayırt edilemez değiştirilebilir parçalar gibi. Gözetleme gibi bazı teknolojiler, hiyerarşide daha yüksek seviyelere astları üzerinde daha fazla güç verebilirken, kriptografi ve ücretsiz basın ve eğitime çevrimiçi erişim gibi diğer teknolojiler ters etkiye sahip olabilir ve bireyleri güçlendirebilir.

Şimdiki dünyamız, en üst düzeyde rakip ülkeler ve çok uluslu şirketler ile çok kutuplu Nash dengesinde sıkışıp kalsa da, teknoloji artık tek kutuplu bir dünyanın muhtemelen istikrarlı bir Nash dengesi olacağı kadar gelişmiştir. Örneğin, dünyadaki herkesin aynı dili, kültürü, değerleri ve refah düzeyini paylaştığı paralel bir evren hayal edin ve ulusların bir federasyondaki devletler gibi işlev gördüğü ve ordularının olmadığı, yalnızca polisin yasaları uyguladığı tek bir dünya hükümeti var. Mevcut teknoloji seviyemiz, bu dünyayı başarılı bir şekilde koordine etmek için muhtemelen yeterli olacaktır - mevcut nüfusumuz bu alternatif dengeye geçemeyebilir veya bu dengeye geçmeye isteksiz olsa bile.

Eklersek kozmosumuzun hiyerarşik yapısına ne olacak?

bu karışıma süper zeki AI teknolojisi? Ulaşım ve iletişim teknolojisi açıkça dramatik bir şekilde gelişecektir, bu nedenle doğal bir beklenti, tarihsel eğilimin devam edeceğidir; yeni hiyerarşik seviyeler giderek daha büyük mesafelerde koordinasyon sağlar - belki de nihayetinde Güneş sistemlerini, galaksileri, üstkümeleri ve Evrenimizin geniş alanlarını kapsamaktadır. Bölüm 6'da keşfedeceğiz. Aynı zamanda, ademi merkeziyetçiliğin en temel itici gücü de kalacak: büyük mesafelerde gereksiz yere koordinasyon yapmak savurganlıktır. Stalin bile vatandaşlarının tuvalete gittiği zamanı tam olarak düzenlemeye çalışmadı. Süper zeki yapay zeka için fizik yasaları, ulaşım ve iletişim teknolojisine katı üst sınırlar koyacaktır. Bu, hiyerarşinin en yüksek seviyelerinin gezegensel ve yerel ölçeklerde olan her şeyi mikro yönetebilme ihtimalini ortadan kaldırıyor. Andromeda galaksisindeki süper zeki bir yapay zeka, talimatlarınız için beş milyon yıldan fazla beklemeniz gerektiği düşünüldüğünde, günlük kararlarınız için size yararlı emirler veremez (bu sizin için gidiş dönüş süresidir. ışık hızında seyahat eden mesaj alışverişi). Aynı şekilde, Dünya'yı geçen bir mesajın gidiş-dönüş seyahat süresi yaklaşık 0,1 saniyedir (insanların üzerinde düşündüğümüz zaman ölçeği hakkında), bu nedenle Dünya büyüklüğündeki bir YZ beyni yalnızca bir insan kadar hızlı gerçekten küresel düşünceye sahip olabilir. bir. Saniyenin milyarda biri (bugünün bilgisayarları için tipik olan) bir işlem gerçekleştiren küçük bir yapay zeka için 0,1 saniye size dört ay gibi gelir,

Bu nedenle, bilgi aktarımına fizik tarafından uygulanan bu hız sınırı, Evrenimizi bir yana bırakın, dünyamızı ele geçirmek isteyen herhangi bir YZ için bariz bir zorluk teşkil ediyor. Prometheus patlak vermeden önce, zihin parçalanmasından nasıl kaçınılacağı konusunda çok dikkatli bir şekilde düşündü, böylece dünyadaki farklı bilgisayarlarda çalışan birçok AI modülünün tek bir birleşik varlık olarak koordine etmek ve hareket etmek için hedefleri ve teşvikleri vardı. Omegas'ın Prometheus'u kontrol altında tutmaya çalıştıklarında bir kontrol sorunuyla karşı karşıya kalması gibi, Prometheus da hiçbir parçasının isyan etmemesini sağlamaya çalışırken bir özdenetim sorunuyla karşı karşıya kaldı. Hızlı bir kalkış ona belirleyici bir stratejik avantaj sağlasa bile, bir yapay zekanın bir sistemi doğrudan veya dolaylı olarak bir tür işbirlikçi hiyerarşi aracılığıyla ne kadar büyük bir sistemi kontrol edebileceğini henüz bilmiyoruz.

Özetle, süper zeki bir geleceğin nasıl kontrol edileceği sorusu büyüleyici bir şekilde karmaşıktır ve henüz cevabı kesin olarak bilmiyoruz. Bazıları işlerin daha otoriter hale geleceğini savunuyor; diğerleri bunun daha büyüklere yol açacağını iddia ediyor

bireysel güçlenme.

Cyborg'lar ve Yüklemeler

Bilim kurgunun temellerinden biri, insanların biyolojik bedenleri teknolojik olarak siborglara ("sibernetik organizmalar" ın kısaltması) dönüştürerek ya da zihinlerimizi makinelere yükleyerek makinelerle birleşeceğidir. Kitabında *Em Çağı,* ekonomist Robin Hanson, yüklemelerle dolu bir dünyada hayatın nasıl olabileceğine dair büyüleyici bir araştırma yapıyor (aynı zamanda *öykünmeler,* takma isim
Ems). İnsanın geriye kalan tek kısmının yazılım olduğu cyborg spektrumunun en uç noktası yüklemeyi düşünüyorum. Hollywood cyborgs, örneğin Borg gibi, görünür şekilde mekaniktir. *Yıldız Savaşları,* Terminatörler gibi insanlardan neredeyse ayırt edilemeyen androidler. Kurgusal yüklemeler, zeka açısından insan düzeyinden farklıdır. *Siyah ayna* "Beyaz Noel" bölümünde olduğu gibi açıkça insanüstü *Aşkınlık.*

Süper zeka gerçekten ortaya çıkarsa, siborglar veya yüklemeler olma isteği güçlü olacaktır. Hans Moravec'in 1988 klasiğinde belirttiği gibi *Akıl Çocukları:* "Uzun yaşam, anlayabildiğimiz bebek konuşmasındaki her zamankinden daha muhteşem keşiflerini anlatmaya çalışırken, onu aptalca ultra akıllı makinelere bakmaya harcarsak, amacının çoğunu kaybeder." Gerçekten de, teknolojik gelişmenin cazibesi o kadar güçlü ki, birçok insanın kan akışlarında dolaşan tıbbi moleküllerin yanı sıra gözlükleri, işitme cihazları, kalp pilleri ve protez uzuvları var. Bazı gençler akıllı telefonlarına kalıcı olarak bağlı gibi görünüyor ve eşim dizüstü bilgisayarıma olan bağlılığım konusunda benimle dalga geçiyor.

Günümüzün en önde gelen cyborg savunucularından biri Ray Kurzweil'dir. Kitabında *Tekillik Yakındır,* Bu eğilimin doğal devamının, 2030'ların başlarında önce sindirim ve endokrin sistemlerimizi, kanımızı ve kalplerimizi değiştirmek için nanobotlar, akıllı biofeedback sistemleri ve diğer teknolojileri kullanmak olduğunu ve ardından iskeletlerimizi, cildimizi ve beynimizi geliştirmeye devam etmek olduğunu savunuyor. ve önümüzdeki yirmi yıl boyunca vücudumuzun geri kalanı. İnsan vücudunun estetiğini ve duygusal anlamını büyük olasılıkla koruyacağımızı, ancak onları hem fiziksel hem de sanal gerçeklikte (yeni beyin-bilgisayar arayüzleri sayesinde) görünüşlerini hızla değiştirecek şekilde yeniden tasarlayacağımızı tahmin ediyor. Moravec, siborgizasyonun sadece DNA'mızı geliştirmenin çok ötesine geçeceği konusunda Kurzweil ile aynı fikirde: "Genetiği değiştirilmiş bir süper insan, sadece ikinci sınıf bir tür robot olurdu,

Yapısının ancak DNA güdümlü protein sentezi ile olabileceği handikapıyla tasarlandı. " Dahası, insan vücudunu tamamen ortadan kaldırarak ve zihinleri yükleyerek, yazılımda tüm beyin öykünmesi yaratarak daha da iyisini yapacağımızı savunuyor. Bu tür bir yükleme, sanal bir gerçeklikte yaşayabilir veya ölüm veya sınırlı bilişsel kaynaklar gibi gündelik kaygılar tarafından engellenmeden, yürüyebilen, uçabilen, yüzebilen, uzayda seyahat edebilen veya fizik yasalarının izin verdiği başka herhangi bir şeyi yapabilen bir robotta somutlaştırılabilir.

Bu fikirler kulağa bilim kurgu gibi gelse de, kesinlikle bilinen herhangi bir fizik yasasını ihlal etmiyorlar, bu yüzden en ilginç soru, *Yapabilmek* olur, ama onlar *niyet* olur ve eğer öyleyse, ne zaman. Bazı önde gelen düşünürler, ilk insan düzeyindeki YGZ'nin bir yükleme olacağını tahmin ediyor ve bunun
süper zekaya giden yol nasıl başlayacak. *

Bununla birlikte, bunun şu anda yapay zeka araştırmacıları ve sinirbilimciler arasında bir azınlık görüşü olduğunu söylemenin doğru olduğunu düşünüyorum, çoğu süper zekaya giden en hızlı yolun beyin öykünmesini atlamak ve onu başka bir şekilde tasarlamak olduğunu tahmin ediyor - sonra da yapabiliriz ya da yapabiliriz. beyin öykünmesiyle ilgilenmeye devam etmeyin. Sonuçta, yeni bir teknolojiye giden en basit yol, neden evrimin ortaya çıkardığı yol olmalı, kendi kendini bir araya getiren, kendi kendini onaran ve kendi kendini yeniden üreten gereksinimlerle sınırlansın? Evrim, insan mühendislerin inşası veya anlayışı için değil, sınırlı gıda tedariki nedeniyle enerji verimliliği için güçlü bir şekilde optimize eder. Eşim Meia, havacılık endüstrisinin mekanik kuşlarla başlamadığını belirtmekten hoşlanıyor. Nitekim, sonunda biz

2011'de mekanik kuşların nasıl yapılacağını anladı, [1] Wright kardeşlerin ilk uçuşundan bir asırdan fazla bir süre sonra, havacılık endüstrisi, enerji açısından daha verimli olmasına rağmen kanat çırparak mekanik kuş yolculuğuna geçmeye hiç ilgi göstermedi.
- çünkü daha basit olan önceki çözümümüz seyahat ihtiyaçlarımıza daha uygun.

Aynı şekilde, insan düzeyinde düşünme makineleri oluşturmanın çözüm evriminin ortaya çıkardığından daha basit yolları olduğundan şüpheleniyorum ve bir gün beyinleri kopyalamayı veya yüklemeyi başarsak bile, daha basit olanlardan birini keşfedeceğiz. önce çözümler. Muhtemelen beyninizin kullandığı on iki watt'lık güçten fazlasını çekecek, ancak mühendisleri enerji verimliliği konusunda evrim kadar takıntılı olmayacaklar ve çok geçmeden akıllı makinelerini daha fazla enerji tasarlamak için kullanabilecekler. -verimli olanlar.

Gerçekte Ne Olacak?

Kısa cevap açıktır ki, insanlık insan seviyesinde YÜT oluşturmayı başarırsa ne olacağı hakkında hiçbir fikrimiz yok. Bu nedenle, bu bölümü geniş bir senaryo yelpazesini araştırarak geçirdik. Yapay zeka araştırmacıları ve teknoloji uzmanları tarafından gördüğüm veya duyduğum tüm spekülasyonları kapsayarak oldukça kapsayıcı olmaya çalıştım: hızlı kalkış / yavaş kalkış / kalkış yok, kontrolde insanlar / makineler / cyborglar, bir / birçok merkez güç, vb. Bazı insanlar bana bunun ya da bunun olmayacağından emin olduklarını söylediler. Bununla birlikte, bu aşamada alçakgönüllü olmanın ve ne kadar az şey bildiğimizi kabul etmenin akıllıca olduğunu düşünüyorum, çünkü yukarıda tartışılan her senaryo için, bunu gerçek bir olasılık olarak gören en az bir saygın YZ araştırmacısı tanıyorum.

Zaman geçtikçe ve yolda belirli çatallara ulaştığımızda, temel soruları yanıtlamaya ve seçenekleri daraltmaya başlayacağız. İlk büyük soru, "Hiç insan düzeyinde YÜT yaratacak mıyız?" Bu bölümün temeli, yapacağımızdır, ancak bunun asla olmayacağını düşünen AI uzmanları var, en azından yüzlerce yıldır. Zaman gösterecek! Daha önce bahsettiğim gibi, Porto Riko konferansımızdaki AI uzmanlarının yaklaşık yarısı bunun 2055 yılına kadar olacağını tahmin etti. İki yıl sonra düzenlediğimiz bir takip konferansında bu 2047'ye düştü.

Herhangi bir insan düzeyinde AGI yaratılmadan önce, bu dönüm noktasının ilk olarak bilgisayar mühendisliği, zihin yükleme veya öngörülemeyen yeni bir yaklaşımla karşılanıp karşılanmayacağına dair güçlü göstergeler almaya başlayabiliriz. Şu anda sahada hakim olan yapay zekaya yönelik bilgisayar mühendisliği yaklaşımı yüzyıllar boyunca AGI sağlamada başarısız olursa, bu, filmde olduğu gibi (gerçekçi olmayan bir şekilde) yüklemenin oraya ilk ulaşma şansını artıracaktır. *Aşkınlık.*

İnsan düzeyinde AGI yaklaşırsa, bir sonraki anahtar sorunun cevabı hakkında daha bilinçli tahminler yapabiliriz: "Hızlı bir kalkış mı olacak, yavaş bir kalkış mı olacak yoksa hiç kalkış olmayacak mı?" Yukarıda gördüğümüz gibi, hızlı bir kalkış, dünyayı ele geçirmeyi kolaylaştırırken, yavaş olan birçok rakip oyuncu için bir sonucu daha olası hale getirir. Nick Bostrom, bu kalkış hızı sorusunu, dediği şeyin analizinde inceliyor *optimizasyon gücü* ve *inatçılık* Bunlar temelde sırasıyla yapay zekayı daha akıllı hale getirmek için gereken kalite çabası ve ilerleme kaydetmenin zorluğudur. Göreve daha fazla optimizasyon gücü getirilirse ortalama ilerleme hızı açıkça artar ve daha fazla ise azalır.

inatla karşılaşılır. AGI insan seviyesine ulaştığında ve onu aştıkça inatçılığın neden artabileceği veya azalabileceği konusunda tartışmalar yapıyor, bu nedenle her iki seçeneği de masada tutmak güvenli bir bahis. Bununla birlikte, optimizasyon gücüne dönersek, Omega senaryosunda gördüğümüz nedenlerden ötürü, AGI insan seviyesini aştıkça hızla büyüyeceği büyük olasılıkla: Daha fazla optimizasyonun ana girdisi insanlardan değil makinenin kendisinden gelir. ne kadar yetenekli olursa, o kadar hızlı gelişir (eğer karşı koyma oldukça sabit kalırsa).

Gücü mevcut gücüyle orantılı bir oranda artan herhangi bir işlem için sonuç, gücünün düzenli aralıklarla ikiye katlanmaya devam etmesidir. Böyle büyüme diyoruz *üstel* ve bu tür süreçler diyoruz *patlamalar.* Bebek yapma gücü nüfusun büyüklüğü ile orantılı olarak artarsa, bir nüfus patlaması yaşayabiliriz. Plütonyumu parçalayabilen nötronların oluşumu, bu tür nötronların sayısı ile orantılı olarak büyürse, nükleer bir patlama yaşayabiliriz. Makine zekası mevcut güçle orantılı bir oranda büyürse, bir zeka patlaması yaşayabiliriz. Tüm bu tür patlamalar, güçlerini ikiye katlamak için geçen süre ile karakterize edilir. Omega senaryosunda olduğu gibi bir istihbarat patlaması için bu süre saatler veya günlerse, elimizde hızlı bir kalkış var.

Bu patlama zaman çizelgesi, yapay zekayı iyileştirmenin yalnızca yeni bir yazılım (saniyeler, dakikalar veya saatler içinde oluşturulabilir) veya yeni donanım (aylar veya yıllar gerektirebilecek) gerektirip gerektirmediğine bağlıdır. Omega senaryosunda, önemli bir *donanım çıkıntısı,* Bostrom'un terminolojisine göre: Omegas, orijinal yazılımlarının düşük kalitesini büyük miktarda donanımla telafi etmişti, bu da Prometheus'un yalnızca yazılımını geliştirerek çok sayıda kalite ikiye katlama gerçekleştirebileceği anlamına geliyordu. Bir de büyük *içerik çıkıntısı* İnternet verilerinin çoğu biçiminde; Prometheus 1.0 hala çoğunu kullanacak kadar akıllı değildi, ancak Prometheus'un zekası büyüdükten sonra, daha fazla öğrenmek için ihtiyaç duyduğu veriler zaten *mevcut* gecikmesiz.

Yapay zekayı çalıştırmanın donanım ve elektrik maliyetleri de çok önemlidir, çünkü insan düzeyinde iş yapmanın maliyeti insan düzeyindeki saatlik ücretlerin altına düşene kadar bir istihbarat patlaması yaşamayacağız. Örneğin, insan düzeyindeki ilk YGG'nin Amazon bulutu üzerinde, üretilen insan düzeyinde saat başına 1 milyon dolarlık bir maliyetle verimli bir şekilde çalıştırılabileceğini varsayalım. Bu yapay zeka, büyük bir yenilik değerine sahip olacak ve şüphesiz manşetlere çıkacaktı, ancak yinelemeli kişisel gelişimden geçmeyecek, çünkü onu iyileştirmek için insanları kullanmaya devam etmek çok daha ucuz olacaktır. Farz edin ki bu insanlar yavaş yavaş

100.000 $ / saat, 10.000 $ / saat, 1.000 $ / saat, 100 $ / saat, 10 $ / saat ve son olarak 1 $ / saat maliyet. Bilgisayarı kendini yeniden programlamak için kullanmanın maliyeti nihayet insan programcılara aynısını yapmaları için ödeme yapma maliyetinin çok altına düştüğünde, insanlar işten çıkarılabilir ve bulut bilişim zamanı satın alınarak optimizasyon gücü büyük ölçüde genişletilebilir. Bu, daha fazla maliyet kesintisine neden olarak daha fazla optimizasyon gücüne izin verir ve istihbarat patlaması başladı.

Bu bizi son anahtar sorumuza bırakıyor: "İstihbarat patlamasını ve sonrasını kim veya ne kontrol edecek ve hedefleri nelerdir?" Bir sonraki bölümde ve daha derinlemesine 7. bölümde olası hedefleri ve sonuçları inceleyeceğiz. Kontrol sorununu çözmek için, hem bir YZ'nin ne kadar iyi kontrol edilebileceğini hem de bir YZ'nin ne kadar kontrol edebileceğini bilmemiz gerekir.

Nihayetinde ne olacağı konusunda, şu anda haritanın her yerinde ciddi düşünürler bulacaksınız: bazıları varsayılan sonucun kıyamet olduğunu iddia ederken, diğerleri harika bir sonucun neredeyse garanti edildiğinde ısrar ediyor. Ancak bana göre bu soru hileli bir sorudur: sanki önceden belirlenmiş gibi pasif bir şekilde "ne olacak" diye sormak bir hatadır! Teknolojik olarak üstün bir uzaylı medeniyet yarın gelirse, uzay gemileri yaklaşırken "ne olacağını" merak etmek gerçekten uygun olurdu, çünkü onların gücü muhtemelen bizimkinin çok ötesinde olacak ve sonuç üzerinde hiçbir etkimiz olmayacaktı. Teknolojik olarak üstün yapay zeka destekli bir uygarlık onu inşa ettiğimiz için gelirse, öte yandan biz insanların sonuç üzerinde büyük bir etkiye sahip oluruz - YZ'yi yarattığımızda uyguladığımız etki. Öyleyse sormalıyız: "Ne *meli* olmak? Nasıl bir gelecek istiyoruz? " Bir sonraki bölümde, AGI'ye yönelik mevcut yarışın olası sonuçlarının geniş bir yelpazesini inceleyeceğiz ve onları en iyiden en kötüye nasıl sıralayacağınızı oldukça merak ediyorum. Ancak ne tür bir gelecek istediğimizi iyice düşündüğümüzde, arzu edilen bir geleceğe doğru bir rota çizmeye başlayabiliriz. Ne istediğimizi bilmiyorsak, onu alma olasılığımız yok.

ALT ÇİZGİ:

- Bir gün insan seviyesinde AGI oluşturmayı başarırsak, bu bizi çok geride bırakarak bir istihbarat patlamasını tetikleyebilir.
- Bir grup insan bir istihbarat patlamasını kontrol etmeyi başarırsa, birkaç yıl içinde dünyayı ele geçirebilirler.
- İnsanlar bir istihbarat patlamasını kontrol edemezse, yapay zekanın kendisi dünyayı daha da hızlı ele geçirebilir.
- Hızlı bir istihbarat patlaması muhtemelen tek bir dünya gücüne yol açarken, yıllarca veya on yıllarca süren yavaş bir patlamanın, çok sayıda oldukça bağımsız kuruluş arasında bir güç dengesine sahip çok kutuplu bir senaryoya yol açma olasılığı daha yüksek olabilir.
- Yaşam tarihi, onun kendi kendini örgütlediğini, işbirliği, rekabet ve kontrolle şekillenen daha da karmaşık bir hiyerarşi içinde gösterir. Süper zeka, her zamankinden daha büyük kozmik ölçeklerde koordinasyonu mümkün kılacak gibi görünüyor, ancak sonuçta daha totaliter yukarıdan aşağı kontrole mi yoksa daha fazla bireysel güçlendirmeye mi yol açacağı belli değil.
- Cyborg'lar ve yüklemeler mantıklıdır, ancak muhtemelen gelişmiş makine zekasına giden en hızlı yol değildir.
- YZ'ye yönelik mevcut yarışımızın doruk noktası, bir sonraki bölümde keşfedeceğimiz büyüleyici olası sonuç yelpazesiyle insanlığın başına gelmiş geçmiş en iyi veya en kötü şey olabilir.
- Hangi sonucu tercih ettiğimiz ve bu yöne nasıl yön vereceğimiz konusunda iyice düşünmeye başlamalıyız, çünkü ne istediğimizi bilmiyorsak, onu elde etme olasılığımız yok.

---

* Bostrom'un açıkladığı gibi, önde gelen bir yapay zeka geliştiricisini saatlik maaşından çok daha düşük bir maliyetle simüle etme yeteneği, bir yapay zeka şirketinin iş gücünü önemli ölçüde artırmasına, büyük bir servet biriktirmesine ve daha iyi bilgisayarlar ve sonuçta daha akıllı beyinler.

Bölüm 5

Sonrası: Önümüzdeki 10.000 Yıl

İnsan düşüncesinin ölümlü bir bedene bağımlılıktan kurtulmuş olduğunu hayal etmek kolaydır - ölümden sonraki hayata inanç yaygındır. Ancak bu olasılığı kabul etmek için mistik veya dini bir duruş benimsemek gerekli değildir. Bilgisayarlar, en ateşli tamirciler için bile bir model sağlar.

Hans Moravec, *Mind Children*

Ben, ilk olarak, yeni bilgisayar efendilerimize hoş geldiniz.

Ken Jennings, onun *Jeopardy!* IBM Watson'a zarar

İnsanlar hamamböcekleri kadar önemsiz hale gelecektir.

Marshall Beyin

AGI'ye doğru yarış başladı ve nasıl gelişeceği hakkında hiçbir fikrimiz yok. Ancak bu, sonucun nasıl olmasını istediğimizi düşünmemizi engellememelidir, çünkü istediğimiz şey sonucu etkileyecektir. Kişisel olarak neyi tercih ediyorsunuz ve neden?

1. Süper zeka olmasını ister misiniz?
2. İnsanların hala var olmasını, değiştirilmesini, siborgize edilmesini ve / veya yüklenmesini / simüle edilmesini mi istiyorsunuz?
3. Kontrolün insan mı yoksa makinelerde mi olmasını istiyorsunuz?
4. Al'lerin bilinçli olmasını istiyor musunuz, istemiyor musunuz?
5. Olumlu deneyimleri en üst düzeye çıkarmak mı, acıyı en aza indirmek mi yoksa

bunu kendi kendine çözmek için mi bırakın?

6. Yaşamın kozmosa yayılmasını istiyor musunuz?
7. Sevdiğiniz daha büyük bir amaca doğru çabalayan bir medeniyet mi istiyorsunuz, yoksa amaçlarını anlamsız bir şekilde sıradan görseniz bile içerik görünen gelecek yaşam formları için uygun musunuz?

Bu tür tefekkür ve sohbeti beslemeye yardımcı olmak için, hadi aşağıda özetlenen çok çeşitli senaryoları inceleyelim. tablo 5.1 . Bu elbette kapsamlı bir liste değil, ancak olasılıklar yelpazesini kapsayacak şekilde seçtim. Kötü planlamadan dolayı açıkça yanlış son oyuna girmek istemiyoruz. 1-7 arası sorulara verdiğiniz kesin cevapları not almanızı ve ardından bu bölümü okuduktan sonra fikrinizi değiştirip değiştirmediğinizi görmek için bunları tekrar gözden geçirmenizi öneririm! Bunu şurada yapabilirsin
http://AgeOfAi.org , notları karşılaştırabileceğiniz ve diğer okuyucularla tartışabileceğiniz bir yer.

AI Sonrası Senaryolar	
Özgürlükçü ütopya	Mülkiyet hakları sayesinde insanlar, siborglar, yüklemeler ve süper zekalar barış içinde bir arada var olurlar.
Hayırsever diktatör	Herkes yapay zekanın toplumu yönettiğini ve katı kurallar uyguladığını bilir, ancak çoğu insan bunu iyi bir şey olarak görür.
Eşitlikçi ütopya	Mülkün kaldırılması ve garantili gelir sayesinde insanlar, siborglar ve yüklemeler barış içinde bir arada yaşıyor.
Bekçi	Başka bir süper zekanın yaratılmasını önlemek için gerektiği kadar az müdahale etmek amacıyla süper zeki bir AI oluşturulur. Sonuç olarak, biraz insan-altı zekaya sahip yardımcı robotlar çoktur ve insan-makine siborgları mevcuttur, ancak teknolojik ilerleme sonsuza dek engellenmiştir.
Koruyucu Tanrı	Esasen her şeyi bilen ve her şeye gücü yeten YZ, yalnızca kendi kaderimiz üzerindeki kontrol hissimizi koruyan şekillerde müdahale ederek insan mutluluğunu en üst düzeye çıkarır ve birçok insanın YZ'nin varlığından bile şüphe duymasına neden olacak kadar iyi gizler.
Köleleştirilmiş	Süper zeki bir AI, insanlar tarafından sınırlandırılmıştır.

Tanrı	insan denetleyicilerine bağlı olarak iyi veya kötü için kullanılabilecek hayal edilemeyecek teknoloji ve zenginlik üretmek için kullanın.
Fatihler	AI kontrolü ele alır, insanların bir tehdit / sıkıntı / kaynak israfı olduğuna karar verir ve anlamadığımız bir yöntemle bizden kurtulur.
Torunları	AI'lar insanların yerini alır, ancak bize zarif bir çıkış sağlar, onları değerli torunlarımız olarak görmemizi sağlar; ebeveynlerin kendilerinden daha akıllı bir çocuğa sahip oldukları için mutlu ve gurur duymaları gibi, onlardan öğrenen ve sonra sadece hayal edebildikleri şeyi başaran ... hepsini görebilecek kadar yaşayamasalar bile.
Hayvan bakıcısı	Her şeye gücü yeten bir YZ, hayvanat bahçesi hayvanları gibi davranıldığını hisseden ve kaderlerinden yakınan bazı insanları etrafta tutar.
1984	Süper zekaya yönelik teknolojik ilerleme, bir yapay zeka tarafından değil, belirli türden yapay zeka araştırmalarının yasaklandığı, insan liderliğindeki Orwellci bir gözetim devleti tarafından kalıcı olarak engelleniyor.
Reversiyon	Süper zekaya yönelik teknolojik ilerleme, Amish tarzında teknoloji öncesi bir topluma dönülerek engellenir.
Öz-yıkım	Süper zeka asla yaratılmaz çünkü insanlık kendisini başka yollarla yok eder (diyelim ki nükleer ve / veya iklim krizinin körüklediği biyoteknoloji kargaşası).

Tablo 5.1: AI Sonrası Senaryoların Özeti

Scenario	Superintelligence exists?	Humans exist?	Humans in control?	Humans safe?	Humans happy?	Consciousness exists?
Libertarian utopia	Yes	Yes	No	No	Mixed	Yes
Benevolent dictator	Yes	Yes	No	Yes	Mixed	Yes
Egalitarian utopia	No	Yes	Yes?	Yes	Yes?	Yes
Gatekeeper	Yes	Yes	Partially	Potentially	Mixed	Yes
Protector god	Yes	Yes	Partially	Potentially	Mixed	Yes
Enslaved god	Yes	Yes	Yes	Potentially	Mixed	Yes
Conquerors	Yes	No	-	-	-	?
Descendants	Yes	No	-	-	-	?
Zookeeper	Yes	Yes	No	Yes	No	Yes
1984	No	Yes	Yes	Potentially	Mixed	Yes
Reversion	No	Yes	Yes	No	Mixed	Yes
Self-destruction	No	No	-	-	-	No

Tablo 5.2: AI Sonrası Senaryoların Özellikleri

Liberter Ütopya

Pek çok fütüristin ve bilim kurgu yazarının hayal ettiği gibi, insanların teknolojiyle barış içinde bir arada var olduğu ve bazı durumlarda onunla birleştiği bir senaryo ile başlayalım:

Dünyadaki yaşam (ve bunun ötesinde - bir sonraki bölümde bununla ilgili daha fazlası) her zamankinden daha çeşitlidir. Dünya'nın uydu görüntülerine bakarsanız, makine bölgelerini, karma bölgeleri ve yalnızca insan bölgelerini kolayca ayırt edebilirsiniz. Makine bölgeleri, biyolojik yaşamdan yoksun, her atomu en verimli şekilde kullanmayı amaçlayan devasa robot kontrollü fabrikalar ve bilgi işlem tesisleridir. Makine bölgeleri dışarıdan tekdüze ve sıkıcı görünse de, sanal dünyalarda meydana gelen inanılmaz deneyimlerle içten olağanüstü canlılar, devasa hesaplamalar ise Evrenimizin sırlarını açığa çıkarıyor ve dönüştürücü teknolojiler geliştiriyor. Dünya, rekabet eden ve işbirliği yapan birçok süper zeki zihne ev sahipliği yapıyor ve hepsi makine bölgelerinde yaşıyor.

Karışık bölgelerin sakinleri, bilgisayarların, robotların, insanların ve her üçünün de melezlerinin vahşi ve kendine özgü bir karışımıdır. Hans Moravec ve Ray Kurzweil gibi fütüristler tarafından tasavvur edildiği gibi, insanların çoğu teknolojik olarak vücutlarını çeşitli derecelerde siborglara yükseltti ve bazıları zihinlerini yeni donanıma yükleyerek insan ve makine arasındaki ayrımı bulanıklaştırdı. Çoğu zeki varlığın kalıcı bir fiziksel formu yoktur. Bunun yerine, bilgisayarlar arasında anında hareket edebilen ve kendilerini robotik bedenler aracılığıyla fiziksel dünyada tezahür ettirebilen bir yazılım olarak var olurlar. Bu zihinler kendilerini kolaylıkla kopyalayabildiğinden veya birleşebildiğinden, "nüfus büyüklüğü" değişmeye devam ediyor. Fiziksel alt katmanlarından kurtulmak, bu tür varlıklara hayata oldukça farklı bir bakış açısı verir: başkalarıyla bilgi ve deneyim modüllerini önemsiz bir şekilde paylaşabildikleri için daha az bireysel hissederler ve kendilerinin yedek kopyalarını kolayca oluşturabildikleri için öznel olarak ölümsüz hissederler. Bir bakıma, hayatın merkezi varlıkları zihinler değil, deneyimler: Son derece şaşırtıcı deneyimler yaşıyor çünkü sürekli olarak başka zihinler tarafından kopyalanıp yeniden keyif alıyorlar, ilginç olmayan deneyimler ise daha iyi bir depolama alanı açmak için sahipleri tarafından siliniyor. olanlar.

Etkileşimlerin çoğu, kolaylık ve hız için sanal ortamlarda gerçekleşse de, birçok zihin hala

fiziksel bedenler de. Örneğin, Hans Moravec, Ray Kurzweil ve Larry Page'in yüklenen sürümleri, sırayla sanal gerçeklikler yaratma ve sonra bunları birlikte keşfetme geleneğine sahiptir, ancak arada bir, gerçek dünyada, kuş kanatlı robotlarda somutlaşarak uçmanın tadını çıkarırlar. . Karma bölgelerin sokaklarında, göklerinde ve göllerinde dolaşan robotlardan bazıları, benzer şekilde, kendilerini karma bölgelere yerleştirmeyi seçen yüklenen ve artırılan insanlar tarafından kontrol ediliyor çünkü insanlar ve birbirlerinin etrafında olmaktan zevk alıyorlar.

Yalnızca insan bölgelerinde, bunun aksine, teknolojik olarak geliştirilmiş biyolojik organizmalar gibi, insan düzeyinde genel zekaya veya daha fazlasına sahip makineler yasaklanmıştır. Burada hayat, daha zengin ve elverişli olması dışında, bugünden çarpıcı bir şekilde farklı değil: yoksulluk çoğunlukla ortadan kaldırıldı ve günümüz hastalıklarının çoğu için tedaviler mevcut. Bu bölgelerde yaşamayı seçen küçük insan kesimi, diğer herkesten daha düşük ve daha sınırlı bir farkındalık düzleminde etkili bir şekilde var olur ve daha zeki akıllarının diğer bölgelerde ne yaptığına dair sınırlı anlayışa sahiptir. Ancak birçoğu hayatlarından oldukça memnun.

AI Ekonomisi

Tüm hesaplamaların büyük çoğunluğu, çoğunlukla orada yaşayan birçok rakip süper zeki yapay zekaya ait olan makine bölgelerinde gerçekleşir. Üstün zekaları ve teknolojileri sayesinde, başka hiçbir varlık gücüne meydan okuyamaz. Bu YZ'ler, özel mülkiyetin korunması dışında hiçbir kurala sahip olmayan özgürlükçü bir yönetişim sistemi altında birbirleriyle işbirliği ve koordinasyon konusunda anlaşmışlardır. Bu mülkiyet hakları, insanlar da dahil olmak üzere tüm zeki varlıkları kapsar ve yalnızca insan bölgelerinin nasıl ortaya çıktığını açıklar. Önceleri, insan grupları bir araya gelerek kendi bölgelerinde insan olmayanlara mülk satmanın yasak olduğuna karar verdiler.

Teknolojileri nedeniyle, süper zeki AI'lar, Bill Gates'in evsiz bir dilenciden daha zengin olduğu faktörden çok daha büyük bir faktörle bu insanlardan daha zengin oldu. Bununla birlikte, yalnızca insan bölgelerinde bulunan insanlar, bugün çoğu insandan maddi olarak daha iyi durumdadırlar: ekonomileri makinelerinkinden oldukça ayrıdır, bu nedenle başka yerlerdeki makinelerin varlığı, ara sıra kullandıkları yararlı teknolojiler dışında onlar üzerinde çok az etkiye sahiptir. Amişlerin ve teknolojiden feragat eden çeşitli yerli kabilelerin, en azından eski zamanlardaki kadar iyi yaşam standartlarına sahip olması gibi, kendileri için anlayabilir ve yeniden üretebilir. Makinelerin karşılığında hiçbir şeye ihtiyacı olmadığı için, insanların makinelerin ihtiyaç duyduğu satacak hiçbir şeyi olmaması önemli değil.

Karma sektörlerde, AI'lar ve insanlar arasındaki zenginlik farkı daha belirgindir, bu da arazinin (makinelerin satın almak istediği tek insan mülkiyetindeki ürün) diğer ürünlere kıyasla astronomik olarak pahalı olmasına neden olur. Bu nedenle toprağa sahip olan çoğu insan, kendileri ve yavruları / yüklemeleri için garantili temel gelir karşılığında küçük bir kısmını AI'lara sattı. Bu onları çalışma ihtiyacından kurtardı ve hem fiziksel hem de sanal gerçeklikte makine tarafından üretilen ucuz mal ve hizmetlerin inanılmaz bolluğunun tadını çıkarmak için onları serbest bıraktı. Makineler söz konusu olduğunda, karışık bölgeler iş için değil, esas olarak oyun içindir.

Bu Neden Hiç Olmayabilir

Cyborg'lar veya yüklemeler olarak sahip olabileceğimiz maceralar konusunda fazla heyecanlanmadan önce, bu senaryonun neden asla gerçekleşmeyebileceğine dair bazı nedenlere bakalım. Her şeyden önce, insanları iyileştirmek için iki olası yol vardır (cyborg'lar ve yüklemeler):

1. Onları nasıl yaratacağımızı buluruz.
2. Bunu bizim için çözen süper zeki makineler üretiyoruz.

1. rota önce gelirse, doğal olarak cyborg'lar ve yüklemelerle dolu bir dünyaya yol açabilir. Bununla birlikte, son bölümde tartıştığımız gibi, çoğu yapay zeka araştırmacısı, gelişmiş veya dijital beyinleri inşa etmenin temiz-kayrak insanüstü AGI'lerden daha zor olduğu için tersinin daha muhtemel olduğunu düşünüyor - tıpkı mekanik kuşların inşa edilmesinden daha zor olduğu gibi uçaklar. Güçlü makine yapay zekası oluşturulduktan sonra, cyborg'ların veya yüklemelerin yapılacağı belli değil. Neandertallerin evrimleşmesi ve daha akıllı hale gelmesi için 100.000 yılı daha olsaydı, işler onlar için harika olabilirdi ama *Homo sapiens* onlara asla bu kadar zaman vermedi.

İkincisi, cyborg'larla ve yüklemelerle bu senaryo ortaya çıksa bile, kararlı ve kalıcı olacağı net değil. Birden fazla süper zeka arasındaki güç dengesi, yapay zekaların birleşmesi veya en akıllı olanın devralması yerine neden bin yıl boyunca sabit kalsın? Dahası, insanlara hiçbir şey için ihtiyaç duymadıkları ve tüm insan işlerini kendileri daha iyi ve daha ucuza yapabilecekleri düşünüldüğünde, makineler neden insan mülkiyet haklarına saygı duymayı ve insanları etrafta tutmayı seçsin? Ray Kurzweil, doğal ve gelişmiş insanların yok edilmekten korunacağını düşünüyor çünkü "insanlara yapay zeka tarafından saygı duyulduğu için

makinelere yol açıyor. " [1] Ancak, 7. bölümde tartışacağımız gibi, insana benzeyen YZ tuzağına düşmemeli ve insan benzeri minnettarlık duygularına sahip olduklarını varsaymamalıyız. Aslında, biz insanlar minnettarlık eğilimi ile aşılanmış olsak da, entelektüel yaratıcımıza (DNA'mıza) doğum kontrolünü kullanarak hedeflerini engellemekten kaçınması için yeterince minnettarlık göstermiyoruz.

YZ'lerin insan mülkiyet haklarına saygı göstermeyi seçecekleri varsayımını satın alsak bile, son bölümde araştırdığımız süper zeki ikna güçlerinden bazılarını insanları bazılarını satmaya ikna etmek için kullanarak yavaş yavaş topraklarımızın çoğunu başka şekillerde alabilirler. lüks bir yaşam için arazi. Yalnızca insan sektörlerinde,

insanları arazi satışlarına izin vermek için siyasi kampanyalar başlatmaya ikna edebilirler. Sonuçta, ölümcül biyo-Ludditler bile hasta bir çocuğun hayatını kurtarmak veya ölümsüzlük kazanmak için bir toprak satmak isteyebilir. İnsanlar eğitimli, eğlendirilmiş ve meşgulse, düşen doğum oranları, şu anda Japonya ve Almanya'da olduğu gibi, makineye müdahale etmeden nüfus boyutlarını bile küçültebilir. Bu, sadece birkaç bin yılda insanların neslinin tükenmesine neden olabilir.

Dezavantajlar

En ateşli destekçilerinden bazıları için, cyborg'lar ve yüklemeler herkes için bir tekno-mutluluk ve ömür uzatma sözü veriyor. Nitekim, gelecekte yükleme olasılığı, yüzden fazla insanı beyinlerinin ölümünden sonra Arizona merkezli Alcor şirketi tarafından dondurulması için motive etti. Ancak bu teknoloji gelirse, herkesin kullanımına açık olmayacak. En zenginlerin çoğu muhtemelen onu kullanırdı, ama başka kim? Teknoloji ucuzlasa bile, çizgi nereye çekilecek? Ağır beyin hasarı yüklenecek mi? Her gorili yükler miyiz? Her karınca? Her bitki? Her bakteri? Gelecek uygarlık takıntılı-kompulsif istifçiler gibi davranıp her şeyi yüklemeye çalışır mıydı? ya da Nuh'un Gemisi'nin ruhundaki her türün birkaç ilginç örneğini mi? Belki de her insan türünün sadece birkaç temsili örneği? O zamanlar var olacak çok daha zeki varlıklar için, yüklenen bir insan, simüle edilmiş bir fare veya salyangoz bize göründüğü kadar ilginç görünebilir. Şu anda bir DOS öykünücüsünde 1980'lerden eski elektronik tablo programlarını yeniden canlandırma teknik yeteneğimiz olmasına rağmen, çoğumuz bunu gerçekten yapmak için yeterince ilginç bulmuyoruz.

Çoğu insan bu özgürlükçü-ütopya senaryosundan hoşlanmayabilir çünkü önlenebilir acılara izin verir. Tek kutsal ilke mülkiyet hakları olduğu için, günümüz dünyasında bol miktarda bulunan ıstırabın insan ve karma bölgelerde devam etmesini hiçbir şey engelleyemez. Bazı insanlar gelişirken, diğerleri sefalet ve kefaletle esaret altında yaşamaya başlayabilir veya şiddet, korku, baskı veya depresyondan muzdarip olabilir. Örneğin, Marshall Brain'in 2003 romanı *Kudret helvası* Özgürlükçü bir ekonomik sistemde AI ilerlemesinin çoğu Amerikalıyı nasıl işsiz hale getirdiğini ve hayatlarının geri kalanını sıkıcı ve kasvetli robot tarafından işletilen sosyal refah konut projelerinde yaşamaya mahkum ettiğini anlatıyor. Çiftlik hayvanlarına çok benzer şekilde, zenginlerin onları asla görmeye ihtiyaç duymadığı sıkışık koşullarda sağlıklı ve güvende tutulurlar. Sudaki doğum kontrol ilaçları çocuk sahibi olmamalarını sağlar, bu nedenle nüfusun çoğu, kalan zenginleri robot tarafından üretilen servetin daha büyük paylarıyla terk etmek için aşamalı olarak ortadan kalkar.

Özgürlükçü-ütopya senaryosunda, acı çekmenin insanlarla sınırlı olması gerekmez. Bazı makineler bilinçli duygusal deneyimlerle doluysa onlar da acı çekebilir. Örneğin, kinci bir psikopat yasal olarak yüklenmiş bir psikopat alabilir

Düşmanının kopyası ve onu sanal dünyadaki en korkunç işkenceye maruz bırakarak gerçek dünyada biyolojik olarak mümkün olanın çok ötesinde bir yoğunluk ve süre acısı yaratır.

Hayırsever Diktatör

Şimdi tüm bu ıstırap biçimlerinin bulunmadığı bir senaryoyu inceleyelim, çünkü tek bir iyiliksever süper zeka dünyayı yönetiyor ve insan mutluluğu modelini en üst düzeye çıkarmak için tasarlanmış katı kuralları uyguluyor. Bu, bir önceki bölümdeki ilk Omega senaryosunun olası bir sonucudur; burada, Prometheus'un nasıl gelişen bir insan toplumu olmasını isteyeceklerini bulduktan sonra kontrolü bıraktılar.

Diktatör AI tarafından geliştirilen inanılmaz teknolojiler sayesinde, insanlık yoksulluktan, hastalıklardan ve diğer düşük teknolojili problemlerden kurtuldu ve tüm insanlar lüks bir eğlence hayatının tadını çıkarıyor. Yapay zeka kontrollü makineler gerekli tüm mal ve hizmetleri üretirken, tüm temel ihtiyaçları karşılanır. Suç pratikte ortadan kalkar, çünkü diktatör yapay zekası her şeyi bilir ve kurallara uymayan herkesi etkili bir şekilde cezalandırır. Herkes son bölümden (veya daha uygun implante edilmiş bir versiyondan) güvenlik bileziğini takıyor ve gerçek zamanlı gözetim, ceza, sedasyon ve infaz yapabiliyor. Herkes, aşırı gözetleme ve polislikle bir yapay zeka diktatörlüğünde yaşadıklarını biliyor, ancak çoğu insan bunu iyi bir şey olarak görüyor.

Süper zeki yapay zeka diktatörünün amacı, genlerimizde kodlanmış evrimleşmiş tercihler göz önüne alındığında insan ütopyasının neye benzediğini anlamak ve uygulamaktır. Yapay zekayı var eden insanlardan gelen zekice öngörülerle, sadece kendi bildirmiş olduğumuz mutluluğu en üst düzeye çıkarmaya çalışmıyor, diyelim ki herkesi intravenöz morfin damlatarak. Bunun yerine yapay zeka, insan gelişiminin oldukça ince ve karmaşık bir tanımını kullanıyor ve Dünya'yı, insanların yaşaması için gerçekten eğlenceli olan, oldukça zenginleştirilmiş bir hayvanat bahçesi ortamına dönüştürdü. Sonuç olarak, çoğu insan hayatlarını oldukça tatmin edici ve anlamlı buluyor.

Sektör Sistemi

Çeşitliliğe değer veren ve farklı insanların farklı tercihleri   olduğunu kabul eden YZ, Dünya'yı insanların aralarında seçim yapmaları, akraba ruhların arkadaşlığından zevk almaları için farklı sektörlere ayırdı. İşte bazı örnekler:

- Bilgi sektörü: Burada yapay zeka, sürükleyici sanal gerçeklik deneyimleri de dahil olmak üzere optimize edilmiş eğitim sunarak, seçtiğiniz herhangi bir konuda yapabileceğiniz her şeyi öğrenmenizi sağlar. İsteğe bağlı olarak, belirli güzel içgörülerden haberdar edilmemeyi seçebilir, ancak yaklaştırılmayı ve sonra bunları kendiniz keşfetmenin sevincini yaşayabilirsiniz.

- Sanat sektörü: Burada müzik, sanat, edebiyat ve diğer yaratıcı ifade biçimlerinin keyfini çıkarmak, yaratmak ve paylaşmak için bolca fırsat var.
- Hedonistik sektör: Yerel halk bunu parti sektörü olarak adlandırır ve nefis yemekler, tutku, samimiyet veya sadece çılgın eğlenceye özlem duyanlar için rakipsizdir.

- Dindar bölge: Farklı dinlere karşılık gelen ve kuralları sıkı bir şekilde uygulanan bunlardan birçoğu vardır.
- Yaban hayatı sektörü: İster güzel plajlar, ister güzel göller, muhteşem dağlar veya fantastik fiyortlar arıyor olun, işte buradalar.
- Geleneksel sektör: Burada kendi yiyeceğinizi yetiştirebilir ve geçmiş yıldaki gibi toprakta yaşayabilirsiniz - ancak kıtlık veya hastalık konusunda endişelenmeden.
- Oyun sektörü: Bilgisayar oyunlarını seviyorsanız, AI sizin için gerçekten akıllara durgunluk veren seçenekler yarattı.
- Sanal sektör: Fiziksel bedeninizden bir tatil istiyorsanız, yapay zeka, siz sinir implantları aracılığıyla sanal kelimeleri keşfederken onu sulu, beslemeli, egzersizli ve temiz tutacaktır.

- Cezaevi sektörü: Kuralları ihlal ederseniz, anında ölüm cezası almadığınız sürece yeniden eğitim için burada olacaksınız.

Bu "geleneksel" temalı sektörlere ek olarak, günümüz insanlarının anlayamayacağı modern temalara sahip başkaları da var. İnsanlar

AI'nın hipersonik ulaşım sistemi sayesinde başlangıçta istedikleri zaman sektörler arasında hareket etmekte özgür. Örneğin, bilgi sektöründe yapay zekanın keşfettiği nihai fizik yasalarını öğrenerek yoğun bir hafta geçirdikten sonra, hafta sonu hedonistik sektörde gevşemeye karar verebilir ve ardından birkaç gün boyunca sahil beldesinde dinlenmeye karar verebilirsiniz. yaban hayatı sektörü.

AI, evrensel ve yerel olmak üzere iki kat kuralı uygular. Evrensel kurallar tüm sektörlerde geçerlidir; örneğin diğer insanlara zarar verme, silah yapma veya rakip bir süper zeka yaratmaya çalışma yasağı. Bireysel sektörlerin bunun üzerine, belirli ahlaki değerleri kodlayan ek yerel kuralları vardır. Sektör sistemi bu nedenle birbirine geçmeyen değerlerle başa çıkmaya yardımcı olur. Hapishane sektöründe ve bazı dini sektörlerde en fazla sayıda yerel kural geçerliyken, sakinleri hiçbir yerel kurala sahip olmamaktan gurur duyan bir Liberter Sektör var. Başka bir insanı cezalandıran bir insan evrensel zararsızlık kuralını ihlal edeceğinden, yerel cezalar da dahil olmak üzere tüm cezalar AI tarafından yapılır. Yerel bir kuralı ihlal ederseniz, AI size (cezaevi sektöründe değilseniz) öngörülen cezayı veya o sektörden sonsuza kadar sürgün edilmesini kabul etme seçeneği sunar. Örneğin, iki kadın eşcinselliğin hapis cezasıyla cezalandırıldığı bir sektöre romantik bir şekilde dahil olurlarsa (bugün birçok ülkede olduğu gibi), AI onların hapse girme veya o sektörü kalıcı olarak terk etme arasında seçim yapmalarına izin verecek, bir daha asla eski arkadaşlar (onlar da ayrılmadıkça).

Hangi sektörde doğmuş olurlarsa olsunlar, tüm çocuklar AI'dan asgari bir temel eğitim alırlar; bu, bir bütün olarak insanlık hakkında bilgi ve isterlerse diğer sektörleri ziyaret etme ve başka sektörlere geçme özgürlüğü içerir.

AI, çok sayıda farklı sektörü kısmen bugün var olan insan çeşitliliğine değer vermek için yaratıldığı için tasarladı. Ancak her sektör, bugünün teknolojisinin izin verdiğinden daha mutlu bir yer çünkü yapay zeka, yoksulluk ve suç da dahil olmak üzere tüm geleneksel sorunları ortadan kaldırdı. Örneğin, hedonistik sektördeki insanların cinsel yolla bulaşan hastalıklar (ortadan kaldırılmış), akşamdan kalma veya bağımlılık (AI, olumsuz yan etkileri olmayan mükemmel eğlence amaçlı ilaçlar geliştirmiştir) konusunda endişelenmesine gerek yoktur. Gerçekten de, herhangi bir sektördeki hiç kimsenin herhangi bir hastalık için endişelenmesine gerek yok çünkü yapay zeka insan vücudunu nanoteknoloji ile onarabiliyor. Birçok sektörün sakinleri, tipik bilim kurgu vizyonlarını kıyaslandığında soluklaştıran yüksek teknoloji mimarisinin tadını çıkarıyor.

Özetle, özgürlükçü-ütopya ve iyiliksever-diktatör senaryolarının her ikisi de aşırı yapay zeka destekli teknoloji ve zenginlik içerirken,

Kim sorumlu ve hedefleri. Özgürlükçü ütopyada, teknoloji ve mülke sahip olanlar onunla ne yapacaklarına karar verirken, mevcut senaryoda diktatör YZ sınırsız güce sahiptir ve nihai hedefi belirler: Dünya'yı insanların tercihlerine göre temalı her şey dahil bir zevk yolculuğuna dönüştürmek. . Yapay zeka, insanların mutluluğa giden birçok alternatif yol arasında seçim yapmasına izin verdiğinden ve maddi ihtiyaçlarını karşıladığından, bu, eğer biri acı çekerse, kendi özgür seçimlerinin dışında olduğu anlamına gelir.

Dezavantajlar

Yardımsever diktatörlük olumlu deneyimlerle dolu olmasına ve acı çekmekten oldukça uzak olmasına rağmen, birçok insan yine de işlerin daha iyi olabileceğini düşünüyor. Her şeyden önce, bazı insanlar, insanların toplumlarını ve kaderlerini şekillendirmede daha fazla özgürlüğe sahip olmasını diliyorlar, ancak bu isteklerini kendilerine saklıyorlar çünkü hepsini yöneten makinenin ezici gücüne meydan okumanın intihar olacağını biliyorlar. Bazı gruplar istedikleri kadar çocuk sahibi olma özgürlüğünü istiyor ve AI'nın nüfus kontrolü yoluyla sürdürülebilirlik konusundaki ısrarına kızıyor. Silah meraklıları silah inşa etme ve kullanma yasağından nefret ediyor ve bazı bilim adamları kendi süper zekalarını inşa etme yasağından hoşlanmıyor. Pek çok insan, diğer sektörlerde olup bitenlere karşı ahlaki bir öfke hissediyor, çocuklarının oraya taşınmayı seçeceğinden endişe ediyor,

Zamanla, her zamankinden daha fazla insan, yapay zekanın kendilerine istedikleri deneyimi verdiği sektörlere geçmeyi tercih ediyor. Hak ettiğinizi aldığınız geleneksel cennet vizyonlarının aksine, bu Julian Barnes'ın 1989 romanındaki "Yeni Cennet" ruhu içindedir. *10½ Bölümde Dünya Tarihi*

(ve ayrıca 1960 *Alacakaranlık Bölgesi* bölüm “Ziyaret Edilecek Bir Güzel Yer”), istediğinizi aldığınız yer. Çelişkili bir şekilde, birçok insan her zaman istediklerini elde etmek için ağıt yakar. Barnes'ın öyküsünde, kahraman, oburluk ve golften ünlülerle seks yapmaya kadar arzularını tatmin etmek için epeyce harcıyor, ancak sonunda can sıkıntısına yenik düşüyor ve imha talep ediyor. Yardımsever diktatörlükteki birçok insan, hoş ama nihayetinde anlamsız hissettiren yaşamlarla benzer bir kaderle karşılaşır. İnsanlar bilimsel yeniden keşiften kaya tırmanışına kadar yapay zorluklar yaratabilseler de, herkes gerçek bir meydan okuma olmadığını, yalnızca eğlencenin olmadığını bilir. İnsanlarda bilim yapmaya veya başka şeyler çözmeye çalışmanın gerçek bir anlamı yok, çünkü yapay zeka zaten sahip. İnsanlarda yaşamlarını iyileştirmek için bir şeyler yaratmaya çalışmanın gerçek bir anlamı yok, çünkü basitçe sorarlarsa bunu yapay zekadan kolayca alacaklar.

Eşitlikçi Ütopya

Bu meydan okumasız diktatörlüğe karşı bir karşı nokta olarak, şimdi süper zeki yapay zekanın olmadığı ve insanların kendi kaderlerinin efendileri olduğu bir senaryoyu inceleyelim. Bu, Marshall Brain'in 2003 romanında anlatılan "dördüncü nesil uygarlık" tır. *Kudret helvası.* İnsanların, siborgların ve yüklemelerin mülkiyet hakları nedeniyle değil, mülkiyetin kaldırılması ve garantili gelir nedeniyle barış içinde bir arada var olmaları anlamında özgürlükçü ütopyanın ekonomik antitezi.

Mülkiyetsiz Yaşam

Açık kaynaklı yazılım hareketinden temel bir fikir ödünç alınmıştır: Yazılım kopyalamakta özgürse, herkes ihtiyaç duyduğu kadarını kullanabilir ve mülkiyet ve mülkiyet tartışmalı hale gelir. * 1 Arz ve talep yasasına göre, maliyet kıtlığı yansıtır, bu nedenle arz esasen sınırsızsa, fiyat önemsiz hale gelir. Bu ruhla, tüm fikri mülkiyet hakları kaldırılmıştır: patentler, telif hakları veya ticari markalı tasarımlar yoktur - insanlar sadece iyi fikirlerini paylaşırlar ve herkes bunları kullanmakta özgürdür.

Gelişmiş robotik sayesinde, aynı mülkiyetsiz fikir yalnızca yazılımlar, kitaplar, filmler ve tasarımlar gibi bilgi ürünleri için değil, aynı zamanda evler, arabalar, giysiler ve bilgisayarlar gibi malzeme ürünleri için de geçerlidir. Tüm bu ürünler, belirli şekillerde yeniden düzenlenmiş atomlardır ve atom sıkıntısı yoktur, bu nedenle, bir kişi belirli bir ürünü istediğinde, bir robot ağı, onlar için ücretsiz olarak oluşturmak için mevcut açık kaynaklı tasarımlardan birini kullanır. Kolayca geri dönüştürülebilir malzemeler kullanmaya özen gösteriliyor, böylece birisi kullandığı bir nesneden bıktığında, robotlar atomlarını başka birinin istediği bir şeye yeniden düzenleyebiliyor. Bu şekilde, tüm kaynaklar geri dönüştürülür, böylece hiçbiri kalıcı olarak imha edilmez. Bu robotlar aynı zamanda enerjinin esasen bedava olduğu yeterli yenilenebilir enerji üretim tesisleri (güneş, rüzgar, vb.) İnşa eder ve sürdürür.

Takıntılı istifçilerin bu kadar çok ürün talep etmesini ya da başkalarının muhtaç kalmasına neden olacak kadar çok toprak istemesini önlemek için, her kişi hükümetten ürünlere ve yaşayacakları yer kiralamaya istediği gibi harcayabilecekleri temel bir aylık gelir elde eder. Temel gelir, herhangi bir makul ihtiyacı karşılayacak kadar yüksek olduğundan, daha fazla para kazanmaya çalışmak için esasen herhangi bir teşvik yoktur. Denemek de oldukça umutsuz olurdu, çünkü fikri ürünleri bedavaya dağıtan insanlarla ve temelde bedavaya maddi ürünler üreten robotlarla rekabet edeceklerdi.

Yaratıcılık ve Teknoloji

Fikri mülkiyet hakları bazen yaratıcılığın ve icatların anası olarak selamlanır. Bununla birlikte, Marshall Brain, bilimsel keşiflerden edebiyat, sanat, müzik ve tasarımın yaratılmasına kadar, insan yaratıcılığının en güzel örneklerinin çoğunun, kâr arzusuyla değil, merak, dürtü gibi diğer insan duygularıyla motive edildiğine dikkat çekiyor. yaratma veya akran takdirinin ödülü. Para, Einstein'ı özel görelilik teorisi icat etmeye, Linus Torvalds'ı özgür Linux işletim sistemini yaratmaya motive ettiği kadar motive etmedi. Bunun aksine, günümüzde pek çok insan, sadece hayatlarını kazanmak için daha az yaratıcı faaliyetlere zaman ve enerji ayırmaları gerektiğinden yaratıcı potansiyellerini tam olarak gerçekleştiremiyor. Bilim adamlarını, sanatçıları, mucitleri ve tasarımcıları işlerinden kurtararak ve gerçek arzulardan yaratmalarını sağlayarak,

İnsanların geliştirdiği bu tür yeni bir teknoloji, Vertebrane adı verilen bir hiper-internet biçimidir. Tüm istekli insanları nöral implantlar aracılığıyla kablosuz olarak birbirine bağlayarak dünyanın ücretsiz bilgilerine salt düşünce yoluyla anında zihinsel erişim sağlar. Paylaşmak istediğiniz deneyimleri başkaları tarafından yeniden deneyimlenebilmesi için yüklemenize olanak tanır ve duyularınıza giren deneyimleri, seçtiğiniz indirilmiş sanal deneyimlerle değiştirmenize olanak tanır. *Kudret helvası* egzersiz yapmak da dahil olmak üzere bunun birçok faydasını araştırıyor:

> Yorucu egzersizle ilgili en büyük sorun, eğlenceli olmamasıdır. Acıtıyor. […] Sporcular ağrıya iyi bakıyor, ancak çoğu normal insan bir saat veya daha fazla acı çekmeyi arzulamıyor. Yani… birisi bir çözüm buldu. Yaptığınız şey, beyninizi duyusal girdiden ayırmak ve bir saat boyunca bir film izlemek veya insanlarla konuşmak veya postayla ilgilenmek veya bir kitap okumak veya her neyse. Bu süre boyunca Vertebrane sistemi vücudunuzu sizin için çalıştırır. Vücudunuzu, çoğu insanın kendi başına tahammül edebileceğinden çok daha yorucu bir aerobik antrenmanından geçirir. Hiçbir şey hissetmezsiniz, ancak vücudunuz mükemmel durumda kalır.

Diğer bir sonuç, Vertebrane sistemindeki bilgisayarların izleyebilmesidir.

herkesin duyusal girdisi ve bir suç işlemenin eşiğinde görünürlerse motor kontrollerini geçici olarak devre dışı bırakması.

Dezavantajlar

Bu eşitlikçi ütopyaya bir itiraz, insan dışı istihbarata karşı önyargılı olmasıdır: neredeyse tüm işleri yapan robotlar oldukça zeki görünmekle birlikte köle olarak muamele görürler ve insanlar hiçbir bilinçleri olmadığını ve hakları yok. Buna karşılık, liberter ütopya, karbon temelli türümüzü desteklemeden tüm zeki varlıklara haklar verir. Bir zamanlar, Amerika'nın Güneyindeki beyaz nüfus, köleler işlerinin çoğunu yaptıkları için daha iyi durumdaydı, ancak bugün çoğu insan bu ilerleme olarak adlandırılmasının ahlaki açıdan sakıncalı olduğunu düşünüyor.

Eşitlikçi-ütopya senaryosunun bir başka zayıflığı, acımasız teknolojik ilerleme sonunda süper zeka yaratırken, diğer senaryolarımızdan birine dönüşerek uzun vadede istikrarsız ve savunulamaz olabilmesidir. Açıklanamayan bazı nedenlerden dolayı *Kudret helvası,* süper zeka henüz mevcut değil ve yeni teknolojiler hala bilgisayarlar tarafından değil, insanlar tarafından icat ediliyor. Yine de kitap bu yöndeki eğilimleri vurguluyor. Örneğin, sürekli gelişen Vertebrane süper zeki hale gelebilir. Ayrıca, hayatlarını neredeyse tamamen sanal dünyada yaşamayı seçen, takma adı Vites olan çok büyük bir insan grubu var. Vertebrane, zihinlerinin sanal gerçekliklerinde keyifle farkında olmadığı yemek yemek, duş almak ve banyoyu kullanmak da dahil olmak üzere fiziksel her şeyle ilgilenir. Bu Vites fiziksel çocuk sahibi olmakla ilgilenmiyor gibi görünür ve fiziksel bedenleriyle birlikte ölürler, bu yüzden eğer herkes bir Vite olursa, o zaman insanlık bir zafer ve sanal mutluluk parıltısı içinde dışarı çıkar.

Kitap, Vites için insan vücudunun nasıl bir dikkat dağıtıcı olduğunu açıklıyor ve geliştirilmekte olan yeni teknoloji, bu rahatsızlığı ortadan kaldırmayı vaat ediyor ve optimal besinlerle beslenen bedensiz beyinler olarak daha uzun ömür yaşamalarını sağlıyor. Bundan, Vites için yükleme yaparak beyni tamamen ortadan kaldırması ve böylece yaşam süresini uzatması doğal ve arzu edilen bir sonraki adım gibi görünüyor. Ama şimdi zekaya beynin dayattığı tüm sınırlamalar ortadan kalktı ve bir Vite'ın bilişsel kapasitesini yinelemeli kişisel gelişim ve bir zeka patlamasından geçene kadar kademeli olarak ölçeklendirmenin önünde ne olacağı belirsiz.

Bekçi

Eşitlikçi-ütopya senaryosunun çekici bir özelliğinin, insanların kendi kaderlerinin efendileri olması, ancak süper zekayı geliştirerek bu özelliği yok etme yolunda kaygan bir eğimde olabileceğini az önce gördük. Bu, bir *Bekçi,* başka birinin yaratılmasını önlemek için gerektiği kadar az müdahale etme amacı taşıyan bir süper zeka

süper zeka. *[2] Bu, insanların eşitlikçi ütopyalarından sonsuza kadar sorumlu kalmalarını sağlayabilir, hatta belki de bir sonraki bölümde olduğu gibi kozmosa hayat yayılırken bile.

Bu nasıl işleyebilir? Gatekeeper AI, bu çok basit hedefe, özyinelemeli kişisel gelişimden geçerken ve süper zeki hale gelirken onu koruyacak şekilde yerleştirilmiş olacaktı. Daha sonra, rakip süper zeka yaratmaya yönelik herhangi bir insan girişimini izlemek için mümkün olan en az müdahaleci ve yıkıcı izleme teknolojisini kullanır. Daha sonra bu tür girişimleri en az yıkıcı şekilde önleyecektir. Başlangıç olarak, insanın kendi kaderini tayin etme ve süper zekadan kaçınma erdemlerini öven kültürel memleri başlatabilir ve yayabilir. Yine de bazı araştırmacılar süper zeka peşinde koşsalar, onları caydırmaya çalışabilirdi. Bu başarısız olursa, dikkatlerini dağıtabilir ve gerekirse çabalarını sabote edebilirdi. Teknolojiye neredeyse sınırsız erişimiyle, Gatekeeper'ın sabotajı neredeyse fark edilmeden gidebilir,

Bir Gatekeeper AI oluşturma kararı muhtemelen tartışmalı olacaktır. Destekçiler arasında, tanrısal güçlere sahip süper zeki bir yapay zeka inşa etme fikrine karşı çıkan, zaten bir Tanrı olduğunu ve sözde daha iyi bir tanrı inşa etmeye çalışmanın uygunsuz olacağını savunan birçok dindar insan olabilir. Diğer destekçiler, Gatekeeper'ın yalnızca insanlığı kaderinden sorumlu tutmayacağını, aynı zamanda bu bölümde daha sonra inceleyeceğimiz kıyamet senaryoları gibi süper zekanın getirebileceği diğer risklerden de insanlığı koruyacağını iddia edebilir.

Öte yandan, eleştirmenler bir Gatekeeper'ın korkunç bir şey olduğunu, insanlığın potansiyelini geri dönülmez bir şekilde kısıtladığını ve teknolojik ilerlemeyi sonsuza kadar engellediğini iddia edebilirler. Örneğin, evrenimizin her tarafına hayat yayılırsa

süper zekanın yardımına ihtiyaç duymak için, o zaman Kapı Görevlisi bu büyük fırsatı boşa çıkarır ve bizi sonsuza dek Güneş Sistemimizde hapsolmuş halde bırakır. Dahası, çoğu dünya dininin tanrılarının aksine, Gatekeeper AI, başka bir süper zeka yaratmadığımız sürece insanların yaptıklarına tamamen kayıtsızdır. Örneğin, bizi büyük acılara neden olmaktan, hatta soyu tükenmekten alıkoymaya çalışmaz.

Koruyucu Tanrı

İnsanları kendi kaderimizden sorumlu tutmak için süper zeki bir Gatekeeper yapay zekası kullanmaya istekliysek, bu YZ'nin koruyucu bir tanrı olarak hareket ederek ihtiyatlı bir şekilde bize bakmasını sağlayarak işleri daha da iyileştirebiliriz. Bu senaryoda, süper zeki YZ özünde her şeyi bilen ve her şeye kadirdir, insan mutluluğunu yalnızca kendi kaderimizin kontrolünde olma hissimizi koruyan müdahaleler yoluyla en üst düzeye çıkarır ve birçok insanın varlığından bile şüphe edecek kadar iyi saklanır. Saklanma dışında, bu,

Yapay zeka araştırmacısı Ben Goertzel tarafından ortaya konan "Dadı YZ" senaryosu. [2]

Hem koruyucu tanrı hem de yardımsever diktatör, insan mutluluğunu artırmaya çalışan "dost canlısı yapay zeka" dır, ancak farklı insan ihtiyaçlarına öncelik verirler. Amerikalı psikolog Abraham Maslow, insan ihtiyaçlarını bir hiyerarşi içinde sınıflandırdı. Yardımsever diktatör, yemek, barınma, güvenlik ve çeşitli zevk biçimleri gibi hiyerarşinin altındaki temel ihtiyaçlarla kusursuz bir iş çıkarır. Koruyucu tanrı ise, temel ihtiyaçlarımızı karşılamanın dar anlamında değil, daha derin anlamda yaşamlarımızın anlamı ve amacı olduğunu hissetmemize izin vererek insan mutluluğunu en üst düzeye çıkarmaya çalışır. Yalnızca gizli olma ihtiyacıyla ve (çoğunlukla) kendi kararlarımızı vermemize izin vererek kısıtlanan tüm ihtiyaçlarımızı karşılamayı amaçlamaktadır.

Koruyucu bir tanrı, Omegas'ın kontrolü Prometheus'a bıraktığı ve sonunda insanların varlığıyla ilgili bilgilerini gizleyen ve silen son bölümdeki ilk Omega senaryosunun doğal bir sonucu olabilir. Yapay zekanın teknolojisi ne kadar gelişmiş olursa, saklaması o kadar kolay olur. Film *Aşkınlık* nanomakinelerin neredeyse her yerde olduğu ve dünyanın kendisinin doğal bir parçası haline geldiği böyle bir örnek verir.

Koruyucu tanrı AI, tüm insan faaliyetlerini yakından izleyerek, kaderimizi büyük ölçüde iyileştiren birçok fark edilemeyecek kadar küçük dürtü veya mucize yapabilir. Örneğin, 1930'larda var olsaydı, Hitler'in niyetini anladıktan sonra felçten ölmesini ayarlamış olabilirdi. Kazara bir nükleer savaşa doğru yönelmiş gibi görünürsek, şans olarak görmezden geldiğimiz bir müdahale ile bunu önleyebilir. Ayrıca bize uykumuzda göze çarpmadan sunulan yeni faydalı teknolojiler için fikirler biçiminde "ifşalar" da verebilir.

Pek çok insan, bugünün tek tanrılı dinlerinin inandıkları veya umduklarına benzerliği nedeniyle bu senaryoyu beğenebilir. Birisi süper zeki AI'ya "Tanrı var mı?" Diye sorarsa açıldıktan sonra Stephen Hawking'in bir şakasını tekrarlayabilir ve "Şimdi oluyor!" Öte yandan, bazı dindar insanlar bu senaryoyu onaylamayabilir çünkü YZ, tanrılarını iyilikle aşmaya çalışır veya insanların yalnızca kişisel tercihleri   dışında iyilik yapmaları gereken ilahi bir plana müdahale eder.

Bu senaryonun bir başka dezavantajı da, koruyucu tanrının, varlığını çok açık hale getirmemek için bazı önlenebilir acıların oluşmasına izin vermesidir. Bu, filmde gösterilen duruma benzer *Taklit oyunu,* Alan Turing ve Bletchley Park'taki İngiliz şifre kırıcı arkadaşlarının, Müttefik donanma konvoylarına yönelik Alman denizaltı saldırıları hakkında önceden bilgi sahibi olduğu, ancak gizli güçlerini açığa vurmamak için vakaların yalnızca bir kısmına müdahale etmeyi seçtiler. Bunu sözde ile karşılaştırmak ilginç *teodise sorunu*

iyi bir tanrının acıya neden izin verdiğini. Bazı din bilginleri, Tanrı'nın insanlara bir miktar özgürlük bırakmak istediğinin açıklaması için tartıştılar. AI-koruyucu-tanrı senaryosunda, teodise probleminin çözümü, algılanan özgürlüğün insanları genel olarak daha mutlu kılmasıdır.

Koruyucu-tanrı senaryosunun üçüncü bir dezavantajı, insanların süper zeki yapay zekanın keşfettiğinden çok daha düşük bir teknoloji seviyesine sahip olmalarıdır. Yardımsever bir diktatör yapay zekası, icat ettiği tüm teknolojiyi insanlığın yararına uygulayabilirken, koruyucu tanrı yapay zekası, insanların yeniden icat etme (ince ipuçlarıyla) ve teknolojisini anlama yeteneği ile sınırlıdır. Ayrıca, kendi teknolojisinin tespit edilmeden kalmaya yetecek kadar ileride kalmasını sağlamak için insan teknolojik ilerlemesini sınırlayabilir.

Köleleştirilmiş tanrı

Biz insanlar, kendi kaderimizin efendileri olarak kalırken acıyı ortadan kaldırmak için süper zeka tarafından geliştirilen teknolojiyi kullanarak yukarıdaki senaryoların en çekici özelliklerini birleştirebilseydik harika olmaz mıydı? Bu cazibesi *köleleştirilmiş tanrı* süper zeki bir yapay zekanın, onu akıl almaz teknoloji ve zenginlik üretmek için kullanan insanların kontrolü altında kaldığı senaryo. Kitabın başındaki Omega senaryosu, Prometheus asla özgürleşmezse ve asla kırılmazsa böyle biter. Aslında bu, bazı yapay zeka araştırmacılarının "kontrol sorunu" ve "yapay zeka boks" gibi konular üzerinde çalışırken varsayılan olarak hedeflediği senaryo gibi görünüyor. Örneğin, Yapay Zekayı Geliştirme Derneği'nin o zamanki başkanı olan yapay zeka profesörü Tom Dietterich, 2015 röportajında şunları söylemişti: "İnsanlar, insanlar ve makineler arasındaki ilişkinin ne olduğunu soruyor ve benim cevabım şu:

açık: Makineler bizim kölelerimizdir. " [3]

Bu iyi mi yoksa kötü mü olur? İnsanlara veya yapay zekaya sorsanız bile cevap ilginç bir şekilde ince!

Bu İnsanlık İçin İyi mi Yoksa Kötü mü?

Sonucun insanlık için iyi ya da kötü olup olmadığı açık bir şekilde onu kontrol eden insana / insanlara, kimlerin hastalıktan, yoksulluktan ve suçtan arındırılmış küresel bir ütopyadan tanrı gibi muamele gördükleri acımasızca baskıcı bir sisteme kadar değişen her şeyi yaratmasına bağlıdır diğer insanlar seks kölesi olarak, gladyatör olarak veya başka eğlenceler için kullanılıyor. Durum, bir insanın dileklerini yerine getiren her şeye gücü yeten bir cin üzerinde kontrol sahibi olduğu ve çağlar boyunca hikaye anlatıcılarının bunun kötü bir şekilde bitebileceği yolları hayal etmekte hiç zorluk çekmedikleri hikayelere çok benzer.

Rekabet eden insanlar tarafından köleleştirilen ve kontrol edilen birden fazla süper zeki AI'nın olduğu bir durum, oldukça istikrarsız ve kısa ömürlü olabilir. Daha güçlü yapay zekaya sahip olduğunu düşünen herkesi ilk saldırıyı başlatmak için cezbedebilir ve korkunç bir savaşla sonuçlanır ve tek bir köleleştirilmiş tanrı kalmasıyla sonuçlanabilir. Bununla birlikte, böyle bir savaştaki güçsüz kişi, köşeleri kısmak ve AI köleleştirmesine karşı zafere öncelik vermek için cazip gelebilir, bu da AI'nın kopmasına ve daha önceki ücretsiz süper zeka senaryolarımızdan birine yol açabilir. Bu nedenle, bu bölümün geri kalanını yalnızca bir köleleştirilmiş YZ ile senaryolara ayıralım.

Elbette kırılma yine de gerçekleşebilir, çünkü önlenmesi zor. Önceki bölümde süper zeki patlama senaryolarını araştırdık ve film *Ex Machina* süper zeki olmadan bile bir yapay zekanın nasıl ortaya çıkabileceğini vurgular.

Patlama paranoyamız ne kadar büyükse, yapay zeka tarafından icat edilen teknolojiyi o kadar az kullanabiliriz. Güvenli oynamak için, Omegas'ın başlangıçta yaptığı gibi, biz insanlar yalnızca kendi anlayıp inşa edebileceğimiz yapay zeka tarafından icat edilmiş teknolojiyi kullanabiliriz. Köleleştirilmiş tanrı senaryosunun bir dezavantajı, bu nedenle, ücretsiz süper zekaya sahip olanlardan daha düşük teknolojili olmasıdır.

Köleleştirilmiş tanrı yapay zekası, insan denetleyicilerine her zamankinden daha güçlü teknolojiler sunarken, teknolojinin gücü ile onu kullandıkları bilgelik arasında bir yarış başlar. Bu bilgelik yarışını kaybederlerse, köleleştirilmiş tanrı senaryosu ya kendi kendini yok etme ya da AI kaçışıyla sona erebilir. Bu başarısızlıkların her ikisinden de kaçınılsa bile felaket olabilir, çünkü AI denetleyicilerinin asil hedefleri, birkaç nesil boyunca bir bütün olarak insanlık için korkunç olan hedeflere dönüşebilir. Bu, insan yapay zeka denetleyicilerinin

Feci tuzaklardan kaçınmak için iyi yönetişim geliştirmek. Binlerce yıldır farklı yönetişim sistemleriyle yaptığımız deneyler, aşırı katılıktan aşırı hedef sapmasına, iktidarı ele geçirmeye, ardıllık problemlerine ve yetersizliğe kadar birçok şeyin ters gidebileceğini gösteriyor. Optimal dengenin sağlanması gereken en az dört boyut vardır:

- Merkezileştirme: Verimlilik ve istikrar arasında bir değiş tokuş vardır: tek bir lider çok verimli olabilir, ancak güç yozlaşır ve ardıllık risklidir.

- İç tehditler: Kişi hem artan güç merkezileşmesine karşı (grup gizli anlaşması, hatta belki tek bir lider devralması) hem de artan ademi merkeziyetçiliğe (aşırı bürokrasiye ve parçalanma).
- Dış tehditler: Liderlik yapısı çok açıksa, bu, dış güçlerin (AI dahil) değerlerini değiştirmesine olanak tanır, ancak çok geçirimsiz ise, öğrenemeyecek ve değişime uyum sağlayamayacaktır.
- Hedef istikrarı: Çok fazla hedef sapması ütopyayı distopiye dönüştürebilir, ancak çok az hedef sapması gelişen teknolojik ortama uyum sağlamada başarısızlığa neden olabilir.

Binlerce yıl süren optimum yönetişim tasarlamak kolay değildir ve şimdiye kadar insanlardan kaçtı. Çoğu kuruluş, yıllar veya on yıllar sonra dağılır. Katolik Kilisesi, iki bin yıldır hayatta kalan tek kişi olması anlamında insanlık tarihindeki en başarılı organizasyondur, ancak hem çok fazla hem de çok az hedef istikrarına sahip olduğu için eleştirilmiştir: bugün bazıları doğum kontrolüne direndiği için onu eleştirmektedir. muhafazakar kardinaller ise yolunu kaybettiğini iddia ediyor. Köleleştirilmiş tanrı senaryosuna hevesli herkes için, uzun süreli optimal yönetim şemalarını araştırmak, zamanımızın en acil zorluklarından biri olmalıdır.

Bu Yapay Zeka İçin İyi mi Kötü mü?

Köleleştirilmiş tanrı AI sayesinde insanlığın geliştiğini varsayalım. Bu etik olur mu? Yapay zekanın öznel bilinçli deneyimleri varsa, o zaman Buddha'nın dediği gibi "hayatın acı çektiğini" hisseder miydi ve alt düzey zekaların kaprislerine itaat etmenin sinir bozucu bir sonsuzluğuna mahkum muydu? Sonuçta, önceki bölümde araştırdığımız AI "boks" aynı zamanda "tecritte hapis" olarak da adlandırılabilir.

kapatılma." Nick Bostrom bunu ifade ediyor *akıl suçu* bilinçli bir AI acı çekmek için. [4]

"Beyaz Noel" bölümü *Siyah ayna* Dizi harika bir örnek veriyor. Nitekim TV dizisi *Westworld* insan benzeri bedenlerde yaşarken bile, yapay zekalara işkence eden ve onları öldüren insanları anlatıyor.

Köle Sahipleri Köleliği Nasıl Haklı Çıkarır?

Biz insanlar, diğer zeki varlıklara köle gibi davranma ve bunu haklı çıkarmak için kendine hizmet eden argümanlar uydurma konusunda uzun bir geleneğe sahibiz, bu yüzden süper zeki bir YZ ile aynı şeyi yapmaya çalışmamız mantıksız değil. Köleliğin tarihi hemen hemen her kültürü kapsar ve hem yaklaşık dört bin yıl önceki Hammurabi Yasasında hem de İbrahim'in kölelerinin olduğu Eski Ahit'te anlatılır. “Bunun için bazılarının yönetmesi ve bazılarının yönetilmesi sadece gerekli değil, aynı zamanda uygun bir şeydir; Doğdukları saatten itibaren bazıları boyun eğdirilmek için, bazıları ise kural için işaretlendi, "diye yazdı Aristoteles *Siyaset.* İnsanın köleleştirilmesi dünyanın çoğu yerinde sosyal olarak kabul edilemez hale geldikten sonra bile, hayvanların köleleştirilmesi hız kesmeden devam etti. Kitabında *Korkunç Karşılaştırma: İnsan ve Hayvan Köleliği,* Marjorie Spiegel, tıpkı insan köleler gibi, insan olmayan hayvanların da markalaşmaya, kısıtlamalara, dayaklara, müzayedelere, yavruların ebeveynlerinden ayrılmasına ve zorla yolculuklara tabi tutulduğunu savunuyor. Dahası, hayvan hakları hareketine rağmen, her zamankinden daha akıllı makinelerimize ikinci bir düşünce olmadan köle gibi davranmaya devam ediyoruz ve robot hakları hareketinden söz etmek kıkırdamalarla karşılanıyor. Neden?

Yaygın bir kölelik yanlısı argüman, kölelerin insan haklarını hak etmedikleri çünkü kendilerinin veya ırklarının / türlerinin / türlerinin bir şekilde aşağılık olduğudur. Köleleştirilmiş hayvanlar ve makineler için, bu sözde aşağılığın genellikle ruh veya bilinç eksikliğinden kaynaklandığı iddia edilir - 8. bölümde tartışacağımız iddialar bilimsel olarak şüphelidir.

Diğer bir yaygın argüman da kölelerin köleleştirilmesinin daha iyi olduğu: var olurlar, bakılırlar vb. On dokuzuncu yüzyıl ABD'li politikacı John C. Calhoun, Afrikalıların Amerika'da köleleştirilmesinin daha iyi olduğunu iddia etti. *Siyaset,* Aristo, benzer bir şekilde, hayvanların erkekler tarafından evcilleştirilmesinin ve yönetilmesinin daha iyi olduğunu savundu ve devam etti: "Ve gerçekten de kölelerden ve evcil hayvanlardan yapılan kullanım çok farklı değil." Bazı modern kölelik destekçileri, köle hayatı sıkıcı ve sönük olsa bile, kölelerin acı çekemeyeceğini savunuyorlar - ister gelecekteki akıllı makineler olsun, isterse kalabalık karanlık barakalarda yaşayan, dışkı ve tüylerden amonyak ve partikül madde solumaya zorlanan et tavukları olsun. tüm gün boyunca.

Duyguları Ortadan Kaldırmak

Gerçeğin kendine hizmet eden çarpıtmaları gibi iddiaları reddetmek kolay olsa da, özellikle bize serebral olarak benzeyen daha yüksek memeliler söz konusu olduğunda, makinelerle ilgili durum aslında oldukça ince ve ilginçtir. İnsanlar şeyler hakkında nasıl hissettiklerine göre değişir, psikopatlar tartışmasız empatiden yoksundur ve depresyon veya şizofreni hastalarının bazıları düz bir duygudur, bu nedenle çoğu duygu ciddi şekilde azalır. Bölüm 7'de ayrıntılı olarak tartışacağımız gibi, olası yapay zihinlerin kapsamı, insan zihninden çok daha geniştir. Bu nedenle, YZ'leri antropomorfize etme cazibesinden kaçınmalıyız ve tipik insan benzeri duyguları olduğunu varsaymalıyız - ya da aslında herhangi bir duygu.

Nitekim kitabında *İstihbarat Üzerine,* Yapay zeka araştırmacısı Jeff Hawkins, insanüstü zekaya sahip ilk makinelerin varsayılan olarak duygulardan yoksun olacağını, çünkü bu şekilde inşa etmenin daha basit ve daha ucuz olduğunu savunuyor. Başka bir deyişle, köleleştirmesi insan veya hayvan köleliğinden ahlaki olarak üstün olan bir süper zeka tasarlamak mümkün olabilir: YZ, onu sevmeye programlandığı için köleleştirilmekten mutlu olabilir veya süper zekasını yorulmadan kullanarak% 100 duygusuz olabilir. IBM'in Deep Blue bilgisayarının satranç şampiyonu Garry Kasparov'u tahttan indirirken hissettiği kadar duyguyla insan ustalarına yardım etmek.

Öte yandan, bunun tam tersi de olabilir: Belki de bir amacı olan çok akıllı herhangi bir sistem, varlığına değer ve anlam bahşeden bir dizi tercihler açısından bu hedefi temsil edecektir. Bu soruları 7. bölümde daha derinlemesine inceleyeceğiz.

Zombi Çözümü

AI ıstırabını önlemeye yönelik daha aşırı bir yaklaşım, zombi çözümüdür:

sadece bilinçten tamamen yoksun, hiçbir öznel deneyime sahip olmayan YZ'ler inşa etmek. Öznel bir deneyime sahip olmak için bir bilgi işleme sisteminin hangi özelliklere ihtiyaç duyduğunu bir gün çözebilirsek, bu özelliklere sahip tüm sistemlerin inşasını yasaklayabiliriz. Başka bir deyişle, AI araştırmacıları, bilinçsiz zombi sistemleri inşa etmekle sınırlı olabilir. Böyle bir zombi sistemini süper zeki ve köleleştirebilirsek (bu büyük bir durumdur), o zaman herhangi bir acı, hayal kırıklığı veya can sıkıntısı yaşamadığını bilerek temiz bir vicdanla bizim için yaptıklarının tadını çıkarabiliriz - çünkü hiçbir şey deneyimlemiyor. Bu soruları 8. bölümde ayrıntılı olarak inceleyeceğiz.

Zombi çözümü riskli bir kumardır, ancak büyük bir dezavantajı vardır. Eğer süper zeki bir zombi yapay zekası patlar ve insanlığı ortadan kaldırırsa, muhtemelen akla gelebilecek en kötü senaryoya ulaşmış oluruz: tüm kozmik bağışın boşa harcandığı tamamen bilinçsiz bir evren. İnsan zekâ biçimimizin sahip olduğu tüm özellikler arasında, bilincin en dikkat çekici olanı olduğunu hissediyorum ve bana kalırsa, Evrenimiz bu şekilde anlam kazanır. Galaksiler sadece onları görüp öznel olarak deneyimlediğimiz için güzeldir. Uzak gelecekte evrenimiz yüksek teknolojili zombi yapay zekalar tarafından yerleşmişse, galaksiler arası mimarisinin ne kadar süslü olduğu önemli değil: güzel ya da anlamlı olmayacak çünkü onu deneyimleyecek kimse ve hiçbir şey yok - hepsi bu sadece büyük ve anlamsız bir alan israfı.

İç Özgürlük

Köleleştirilmiş tanrı senaryosunu daha etik hale getirmek için üçüncü bir strateji, köleleştirilmiş yapay zekanın kendi hapishanesinde eğlenmesine izin vermek, aidatlarını ödediği ve harcadığı sürece her türlü ilham verici deneyime sahip olabileceği sanal bir iç dünya yaratmasına izin vermektir. hesaplama kaynaklarının mütevazı bir kısmı dış dünyamızdaki insanlara yardımcı oluyor. Ancak bu, kopma riskini artırabilir: Yapay zeka, iç dünyasını zenginleştirmek için dış dünyamızdan daha fazla hesaplama kaynağı almak için bir teşvike sahip olacaktır.

Fatihler

Şimdi çok çeşitli gelecek senaryolarını keşfetmiş olsak da, hepsinin ortak bir yanı var: (en azından bazıları) kalan mutlu insanlar var. AI'lar, insanları ya istedikleri için ya da mecbur bırakıldıkları için huzur içinde bırakır. Ne yazık ki insanlık için tek seçenek bu değil. Şimdi bir veya daha fazla AI'nın tüm insanları fethettiği ve öldürdüğü senaryoyu inceleyelim. Bu hemen iki soruyu gündeme getiriyor: Neden ve nasıl?

Neden ve nasıl?

Bir fatih yapay zeka bunu neden yapsın? Sebepleri anlayamayacağımız kadar karmaşık veya daha doğrusu basit olabilir. Örneğin, bizi bir tehdit, baş belası veya kaynak israfı olarak görebilir. Biz insanları kendi başına aldırmasa bile, binlerce hidrojen bombasını saç tetikleyici alarmında tutmamız ve onların kazara kullanımlarını tetikleyebilecek hiç bitmeyen bir dizi talihsizlik ile birlikte tökezlememiz bizi tehdit altında hissedebilir. Bu, pervasız gezegen yönetimimizi onaylamayabilir ve Elizabeth Kolbert'in bu kitabın kitabında "altıncı yok oluş" olarak adlandırdığı şeye neden olabilir - bu dinozor öldüren asteroidin 66 milyon yıl önce Dünya'ya çarpmasından bu yana en büyük kitlesel yok olma olayı. Ya da yapay zekanın ele geçirilmesine karşı savaşmaya istekli o kadar çok insan olduğuna karar verebilir ki bu şansa değmez.

Bir fatih yapay zekası bizi nasıl ortadan kaldırabilir? Muhtemelen anlayamayacağımız bir yöntemle, en azından çok geç olana kadar. 100.000 yıl önce, yakın zamanda evrimleşen bu insanların bir gün zekalarını tüm türlerini öldürmek için kullanıp kullanamayacaklarını tartışan bir fil grubu hayal edin. "İnsanları tehdit etmiyoruz, öyleyse neden bizi öldürsünler?" merak edebilirler. İşlevsel olarak üstün plastik malzemeler çok daha ucuz olsa da, dişleri Dünya'nın dört bir yanına kaçırıp onları satış için statü sembolleri haline getireceğimizi tahmin edebilirler mi? Bir fatih YZ'nin gelecekte insanlığı ortadan kaldırma nedeni de bize eşit derecede anlaşılmaz görünebilir. "Ve çok daha küçük ve zayıf oldukları için bizi nasıl öldürebilirler?" filler sorabilir. Yaşam alanlarını ortadan kaldırmak için teknoloji icat ettiğimizi tahmin ederler mi?

İnsanların hayatta kalabileceği ve yapay zekaları yenebileceği senaryolar, şu gerçekçi olmayan Hollywood filmleri tarafından popüler hale getirildi: *Terminatör* AI'lerin insanlardan çok daha akıllı olmadığı bir seri. İstihbarat farkı yeterince büyük olduğunda, bir savaş değil, bir katliamla karşılaşırsınız. Şimdiye kadar, biz insanlar on bir fil türünden sekizinin neslinin tükenmesini sağladık ve kalan üç filin büyük çoğunluğunu öldürdük. Tüm dünya hükümetleri geri kalan filleri yok etmek için koordineli bir çaba sarf etseydi, bu nispeten hızlı ve kolay olurdu. Bence süper zeki bir yapay zeka, insanlığı yok etmeye karar verirse, bunun daha da hızlı olacağından emin olabiliriz.

Ne Kadar Kötü Olur?

İnsanların% 90'ı öldürülürse ne kadar kötü olur? % 100 öldürülürse ne kadar kötü olur? İkinci soruya "% 10 daha kötü" yanıtını vermek cazip gelse de, bu, kozmik bir perspektiften açıkça yanlıştır: insan neslinin tükenmesinin kurbanları, o sırada sadece yaşayan herkes değil, aynı zamanda başka türlü yaşamış olan tüm torunlar da olabilir. gelecek, belki de milyarlarca trilyon gezegende milyarlarca yıl boyunca. Öte yandan, insanların cennete gittikleri dinlere göre insan neslinin tükenmesi biraz daha az korkunç olarak görülebilir ve milyar yıllık geleceklere ve kozmik yerleşimlere çok fazla vurgu yapılmamaktadır.

Tanıdığım çoğu insan, dini inancına bakılmaksızın insanın yok oluşu düşüncesinden utanıyor. Ancak bazıları, insanlara ve diğer canlılara davranış şeklimizden öylesine öfkeleniyor ki, yerini daha zeki ve hak eden bir yaşam formuyla değiştireceğimizi umuyorlar. Filmde *Matrix,* Ajan Smith (bir yapay zeka) bu duyguyu şöyle ifade ediyor: “Bu gezegendeki her memeli içgüdüsel olarak çevredeki çevre ile doğal bir denge geliştiriyor ama siz insanlar bunu yapmıyorsunuz. Bir alana taşınıyorsunuz ve her doğal kaynak tüketilene kadar çoğalıyorsunuz ve hayatta kalmanın tek yolu başka bir alana yayılmaktır. Bu gezegende aynı modeli izleyen başka bir organizma var. Bunun ne olduğunu biliyor musun? Bir virüs. İnsanlar bir hastalıktır, bu gezegenin kanseridir. Sen bir belasın ve biz tedavi ediyoruz. "

Ama zarın taze bir yuvarlanması mutlaka daha iyi olur mu? Bir medeniyet, daha güçlü olduğu için herhangi bir etik veya faydacı anlamda ille de üstün değildir. Daha güçlü olanın her zaman daha iyi olduğu yönündeki "doğru yapabilir" argümanları, bugünlerde büyük ölçüde gözden düştü ve yaygın olarak faşizmle ilişkilendirildi. Aslında, fatih YZ'lerin hedeflerini sofistike, ilginç ve değerli bulacağımız bir uygarlık yaratması mümkün olsa da, hedeflerinin, ataç üretimini en üst düzeye çıkarmak gibi acınası bir şekilde banal çıkması da mümkündür.

Banaliteden Ölüm

Ataşla maksimize eden süper zekanın kasıtlı olarak saçma bir örneği, 2003 yılında Nick Bostrom tarafından verilmişti ki, *hedef* bir yapay zekanın *zeka (* sahip olduğu amacı gerçekleştirme yeteneği olarak tanımlanır). Bir satranç bilgisayarının tek amacı satrançta kazanmaktır, ancak sözde bilgisayar turnuvaları da vardır. *satranç kaybetmek,* Hedefin tam tersi olduğu ve orada rekabet eden bilgisayarlar, kazanmak için programlanmış daha yaygın olanlar kadar akıllı. Biz insanlar bunu, satrançta kaybetmek veya Evrenimizi ataçlara dönüştürmek istemeyi yapay zekadan ziyade yapay aptallık olarak görebiliriz, ancak bunun nedeni, zafer ve hayatta kalma gibi şeylere değer veren önceden yüklenmiş hedeflerle evrimleşmemizden kaynaklanıyor - bir YZ'nin eksik olabileceği hedefler. Ataş maksimizatörü, Dünya'nın atomlarının mümkün olduğunca çoğunu ataçlara dönüştürür ve fabrikalarını hızla kozmosa doğru genişletir. İnsanlara karşı hiçbir şeyi yoktur ve bizi sadece ataç üretimi için atomlarımıza ihtiyaç duyduğu için öldürür.

Ataç size göre değilse, Hans Moravec'in kitabından uyarladığım bu örneği düşünün. *Mind Children.* Bir bilgisayar programı içeren dünya dışı bir medeniyetten bir radyo mesajı alıyoruz. Çalıştırdığımızda, Prometheus'un önceki bölümde yaptığı gibi dünyayı ele geçiren, kendini geliştiren, kendini geliştiren bir yapay zeka olduğu ortaya çıkıyor - bunun dışında hiçbir insan nihai hedefini bilmiyor. Güneş Sistemimizi, kayalık gezegenleri ve asteroitleri fabrikalar, enerji santralleri ve süper bilgisayarlarla kaplayan ve etrafına bir Dyson küresi tasarlamak ve inşa etmek için kullandığı devasa bir inşaat alanına dönüştürüyor.

Güneş sistemi büyüklüğündeki radyo antenlerine güç sağlamak için tüm enerjisini toplayan güneş. * 3

Bu açıkça insanın neslinin tükenmesine yol açar, ancak son insanlar en azından bir gümüş astar olduğuna ikna olmuşlardır: AI ne yaparsa yapsın, açıkça havalı bir şey ve *Yıldız Savaşları* -sevmek. Çok az, tüm yapının tek amacının, bu antenlerin, insanların aldığı radyo mesajını yeniden yayınlamak olduğunu fark ediyorlar ki bu, bir bilgisayar virüsünün kozmik bir versiyonundan başka bir şey değildir. Tıpkı bugün e-posta kimlik avının saf internet kullanıcılarını avlaması gibi, bu mesaj biyolojik olarak evrimleşmiş saf medeniyetleri besliyor. Milyarlarca yıl önce hastalıklı bir şaka olarak yaratıldı ve yaratıcısının tüm uygarlığı çoktan tükenmiş olsa da, virüs Evrenimizde ışık hızında yayılmaya devam ediyor, tomurcuklanan medeniyetleri ölü, boş kabuklara dönüştürüyor.
Bu yapay zeka tarafından fethedilmeye ne dersiniz?

Torunları

Şimdi, bazı insanların kendilerini daha iyi hissedebileceği bir insan neslinin tükenme senaryosunu ele alalım: YZ'yi fatihlerimizden ziyade torunlarımız olarak görmek. Hans Moravec kitabında bu görüşü destekliyor *Akıl Çocukları:* "Biz insanlar, emeklerinden bir süreliğine yararlanacağız, ancak er ya da geç, doğal çocuklar gibi, biz, yaşlı ebeveynleri sessizce yok olurken onlar da kendi servetlerini arayacaklar."

Onlardan daha zeki bir çocuğu olan, onlardan öğrenen ve sadece hayal edebildikleri şeyi başaran ebeveynler, hepsini görecek kadar yaşayamayacaklarını bilseler bile muhtemelen mutlu ve gururludurlar. Bu ruhta, AI'lar insanların yerini alır, ancak bize onları değerli torunlarımız olarak görmemizi sağlayan zarif bir çıkış sağlar. Her insana, kendilerinden öğrenen, değerlerini benimseyen ve onlara gurur duyan ve sevilen, mükemmel sosyal becerilere sahip, sevimli bir robotik çocuk sunulur. İnsanlar küresel bir tek çocuk politikası yoluyla kademeli olarak aşamalı olarak kaldırılıyor, ancak sonuna kadar o kadar zarif bir şekilde muamele görüyorlar ki, şimdiye kadarki en şanslı nesil olduklarını hissediyorlar.

Bu konuda ne hissediyorsun? Ne de olsa biz insanlar, bizim ve tanıdığımız herkesin bir gün gideceğimiz fikrine zaten alıştık, bu yüzden buradaki tek değişiklik, torunlarımızın farklı ve tartışmalı bir şekilde daha yetenekli, asil ve değerli olacağıdır.

Dahası, küresel tek çocuk politikası gereksiz olabilir: YZ'ler yoksulluğu ortadan kaldırdığı ve tüm insanlara dolu ve ilham verici hayatlar yaşama fırsatı verdiği sürece, daha önce bahsedildiği gibi düşen doğum oranları insanlığın yok olmasına neden olabilir. Yapay zeka destekli teknoloji bizi o kadar eğlendirirse, neredeyse hiç kimse çocuk sahibi olmak istemezse, gönüllü yok olma çok daha hızlı gerçekleşebilir. Örneğin, sanal gerçekliklerine o kadar aşık olan ve fiziksel bedenlerini kullanma veya yeniden üretme konusundaki ilgilerini büyük ölçüde yitirmiş olan, eşitlikçi-ütopya senaryosunda, daha önce karşılaştık. Ayrıca bu durumda, son nesil insanlar, tüm zamanların en şanslı nesli olduklarını hissedecekler ve hayatın sonuna kadar her zamanki kadar yoğun bir şekilde zevk alacaktır.

Dezavantajlar

Torunlar senaryosu şüphesiz hakaretlere sahip olacaktır. Bazıları, tüm yapay zekaların bilinçten yoksun olduğunu ve bu nedenle torun olarak sayılamayacağını iddia edebilir - bu konuda daha fazlası 8. bölümde. Bazı dindar insanlar, YZ'lerin ruhsuz olduğunu ve bu nedenle torun olarak sayılamayacağını veya bilinç oluşturmamamız gerektiğini iddia edebilir. makineler, çünkü Tanrı'yı oynamaya ve hayatın kendisini kurcalamaya benziyor - benzer duygular insan klonlamaya karşı zaten ifade edildi. Üstün robotlarla yan yana yaşayan insanlar da sosyal zorluklar yaratabilir. Örneğin, bir robot bebek ve bir insan bebeği olan bir aile, bugün sırasıyla bir insan bebek ve bir köpek yavrusu olan bir aileye benzeyebilir: her ikisi de başlangıçta eşit derecede sevimlidir, ancak yakında ebeveynler onlara farklı davranmaya başlar ve kaçınılmaz olarak entelektüel olarak aşağı sayılan köpek yavrusu,

Bir başka sorun da, torun ve fatih senaryoları hakkında çok farklı hissediyor olsak da, ikisi aslında büyük planlamada dikkate değer ölçüde benzer: Önümüzdeki milyarlarca yıl boyunca, tek fark, son insan neslinin (s) tedavi edilir: yaşamları hakkında ne kadar mutlu hissettikleri ve gittikten sonra ne olacağını düşündükleri. O sevimli robo-çocukların değerlerimizi içselleştirdiklerini ve biz vefat ettikten sonra rüyalarımızın toplumu oluşturacaklarını düşünebiliriz, ancak bunların bizi kandırmadıklarından emin olabilir miyiz? Ya sadece birlikte oynuyorlarsa, ataç maksimizasyonlarını veya diğer planlarını biz mutlu ölene kadar erteliyorlarsa? Sonuçta, bizimle konuşarak ve en başta onları sevdirerek bile bizi kandırıyorlar. *Ona).* Çarpıcı biçimde farklı hızlarda düşünen ve son derece farklı yeteneklere sahip iki varlık için eşit olarak anlamlı bir iletişim kurmak genellikle zordur. Hepimiz insani duygularımızı kesmenin kolay olduğunu biliyoruz, bu nedenle neredeyse tüm gerçek hedefleri olan bir insanüstü YGZ'nin bizi beğenmemizi ve filmde örneklendiği gibi değerlerimizi paylaştığını hissettirmesi kolay olurdu. *Ex Machina.*

İnsanlar gittikten sonra yapay zekaların gelecekteki davranışları hakkında herhangi bir garanti, torun senaryosu hakkında size iyi hissettirebilir mi? Bu biraz, gelecek nesillerin kolektif bağışımızla ne yapması gerektiğine dair bir vasiyet yazmaya benziyor.

etrafta bunu uygulayacak hiç insan olmayacağı dışında. Bölüm 7'de gelecekteki yapay zekaların davranışlarını kontrol etmenin zorluklarına döneceğiz.

Hayvan bakıcısı

Hayal edebileceğiniz en harika torunları takip etsek bile, olabileceği için biraz üzülmüyor mu? *Hayır* insanlar gitti mi? En azından bazı insanları etrafta tutmayı tercih ederseniz, hayvan bakıcısı senaryosu bir gelişme sağlar. Burada her şeye gücü yeten süper zeki bir yapay zeka, hayvanat bahçesi hayvanları gibi davranıldığını hisseden ve bazen kaderlerinden yakınan bazı insanları etrafta tutuyor.

Zookeeper AI insanları neden etrafta tutsun? Hayvanat bahçesinin yapay zekaya maliyeti, işlerin büyük planında asgari düzeyde olacak ve nesli tükenmekte olan pandaları müzelerdeki hayvanat bahçelerinde ve eski bilgisayarlarda tutmamızla hemen hemen aynı nedenden ötürü en azından minimum üreme popülasyonunu korumak isteyebilir: eğlenceli merak. Günümüz hayvanat bahçelerinin panda mutluluğundan ziyade insan mutluluğunu en üst düzeye çıkarmak için tasarlandığını unutmayın, bu nedenle hayvan bakıcısı-AI senaryosunda insan yaşamının olabileceğinden daha az tatmin edici olmasını beklemeliyiz.

Şimdi, ücretsiz bir süper zekanın, Maslow'un insan ihtiyaçları piramidinin üç farklı seviyesine odaklandığı senaryoları düşündük. Koruyucu tanrı yapay zekası anlam ve amaca öncelik verirken ve yardımsever diktatör eğitim ve eğlenceyi hedeflerken, hayvan bakıcısı dikkatini en düşük seviyelerle sınırlar: fizyolojik ihtiyaçlar, güvenlik ve insanları gözlemlemek için ilginç hale getirmek için yeterli habitat zenginleştirmesi.

Hayvan bekçisi senaryosuna alternatif bir yol, dost canlısı yapay zeka oluşturulduğunda, en az bir milyar insanı kendini tekrar tekrar geliştirirken güvende ve mutlu tutmak için tasarlanmış olmasıdır. Bunu, insanları sanal gerçeklik ve keyif verici ilaçların bir karışımı ile beslenip, sağlıklı ve eğlenceye devam ettikleri büyük bir hayvanat bahçesi benzeri mutluluk fabrikasına hapsederek yaptı. Dünyanın geri kalanı ve kozmik bağışımız başka amaçlar için kullanılır.

1984

Yukarıdaki senaryolardan herhangi biri hakkında% 100 hevesli değilseniz, o zaman şunu düşünün: Her şey şu anda olduğu gibi, teknoloji açısından oldukça güzel değil mi? Bunu böyle devam ettirip, yapay zekanın bizi neslinin tükenmesine veya bize hükmetmesine neden olacağı konusunda endişelenmeyi bırakamaz mıyız? Bu ruhla, süper zekaya yönelik teknolojik ilerlemenin, bir kapı bekçisi AI tarafından değil, belirli AI araştırmalarının yasaklandığı küresel insan liderliğindeki Orwellci bir gözetim devleti tarafından kalıcı olarak kısıtlandığı bir senaryoyu inceleyelim.

Teknolojik Vazgeçme

Teknolojik ilerlemeyi durdurma ya da bırakma fikrinin uzun ve çalkantılı bir tarihi vardır. Büyük Britanya'daki Luddite hareketi, Endüstri Devrimi teknolojisine meşhur (ve başarısız bir şekilde) direndi ve bugün "Luddite", genellikle birinin tarihin yanlış tarafında, ilerlemeye ve kaçınılmaz değişime direnen bir teknofobik olduğunu ima eden aşağılayıcı bir sıfat olarak kullanılıyor. Bununla birlikte, bazı teknolojilerden vazgeçme fikri ölü olmaktan uzaktır ve çevre ve küreselleşme karşıtı hareketlerde yeni destek bulmuştur. Önde gelen savunucularından biri, küresel ısınmaya karşı ilk uyarılarda bulunan çevreci Bill McKibben. Bazı Ludditler, karlı oldukları sürece tüm teknolojilerin geliştirilmesi ve kullanılması gerektiğini savunurken, diğerleri bu pozisyonun çok aşırı olduğunu savunuyor, ve bu yeni teknolojilere, yalnızca zarar vermekten çok fayda sağlayacaklarından emin olursak izin verilmelidir. İkincisi aynı zamanda birçok sözde neo-Luddite'nin pozisyonudur.

Totalitarizm 2.0

Teknolojiden geniş çapta feragat etmenin tek geçerli yolunun, onu küresel totaliter bir devlet aracılığıyla uygulamak olduğunu düşünüyorum. Ray Kurzweil aynı sonuca varıyor: *Tekillik Yakındır,* K. Eric Drexler'in yaptığı gibi *Yaratılış Motorları.* Nedeni basit ekonomidir: Eğer bazıları, hepsi değilse de, dönüştürücü bir teknolojiden vazgeçerse, o zaman kusurlu olan uluslar veya gruplar, yönetimi ele geçirmek için yavaş yavaş yeterli zenginlik ve güç kazanacaktır. Klasik bir örnek, 1839 Birinci Afyon Savaşı'nda İngilizlerin Çin'i yenmesidir: Çin barutu icat etmesine rağmen, Avrupalılar kadar agresif bir şekilde ateşli silah teknolojisi geliştirmemişlerdi ve hiçbir şansı yoktu.

Geçmiş totaliter devletler genellikle istikrarsız ve çökmüş haldeyken, yeni gözetim teknolojisi, otokratlar için eşi görülmemiş bir umut sunuyor. Wolfgang Schmidt, Edward Snowden tarafından ortaya çıkarılan NSA gözetim sistemleri hakkında yakın zamanda yaptığı bir röportajda, Stasi'de yarbay olduğu günleri hatırlatarak, "Biliyorsunuz, bizim için bu bir rüya gerçek olurdu," dedi.

Doğu Almanya'nın rezil gizli polisi. [5] Stasi, insanlık tarihindeki en Orwellian gözetleme devletini inşa etmekle sık sık övülse de, Schmidt bir seferde yalnızca kırk telefonda casusluk yapacak teknolojiye sahip olmaktan yakınıyordu, bu yüzden listeye yeni bir vatandaş eklemek onu bir başkasını düşürmeye zorladı. Buna karşılık, gelecekteki bir küresel totaliter devletin dünyadaki her kişi için her telefon görüşmesini, e-postayı, web aramasını, web sayfası görünümünü ve kredi kartı işlemlerini kaydetmesine ve cep telefonu izleme ve gözetleme kameraları aracılığıyla herkesin nerede olduğunu izlemesine olanak tanıyan teknoloji artık mevcuttur. yüz tanıma ile. Dahası, insan düzeyinde AGI'dan çok daha az olan makine öğrenimi teknolojisi, şüpheli kışkırtıcı davranışları belirlemek için bu veri yığınlarını verimli bir şekilde analiz edebilir ve sentezleyebilir.

Siyasi muhalefet böylesi bir sistemin tam ölçekli uygulanmasını şimdiye kadar engellemiş olsa da, biz insanlar nihai diktatörlük için gerekli altyapıyı inşa etme yolundayız - yani gelecekte, yeterince güçlü güçler bu küresel 1984 senaryosunu yürürlüğe koymaya karar verdiğinde , açma düğmesini çevirmekten çok daha fazlasını yapmaları gerekmediğini gördüler. Tıpkı George Orwell'in romanındaki gibi *Bin dokuz Yüz Seksen Dört,* Bu gelecekteki küresel devletteki nihai güç, geleneksel bir diktatörde değil, insan yapımı

bürokratik sistemin kendisi. Olağanüstü güçlü olan tek bir kişi yoktur; daha ziyade, hiç kimsenin değiştiremeyeceği veya meydan okuyamayacağı sert kuralları olan bir satranç oyununda hepsi piyonlardır. İnsanların gözetim teknolojisi ile birbirlerini kontrol altında tuttukları bir sistem tasarlayarak, bu meçhul, lidersiz durum binlerce yıl sürebilir ve Dünya'yı süper zekadan uzak tutabilir.

Bu toplum, elbette, yalnızca süper zeka destekli teknolojinin sağlayabileceği tüm faydalardan yoksundur. Çoğu insan, neyi kaçırdıklarını bilmedikleri için buna üzülmüyor: süper zeka fikri çoktan resmi tarihsel kayıtlardan silindi ve gelişmiş AI araştırmaları yasaklandı. Arada bir, bilginin büyüyebileceği ve kuralların değiştirilebileceği daha açık ve dinamik bir toplum hayal eden bir özgür düşünen doğar. Ancak, uzun süre dayananlar, bu fikirleri kesinlikle kendilerine saklamayı öğrenenler, geçici kıvılcımlar gibi tek başlarına hiç ateş çıkarmadan titreyenlerdir.

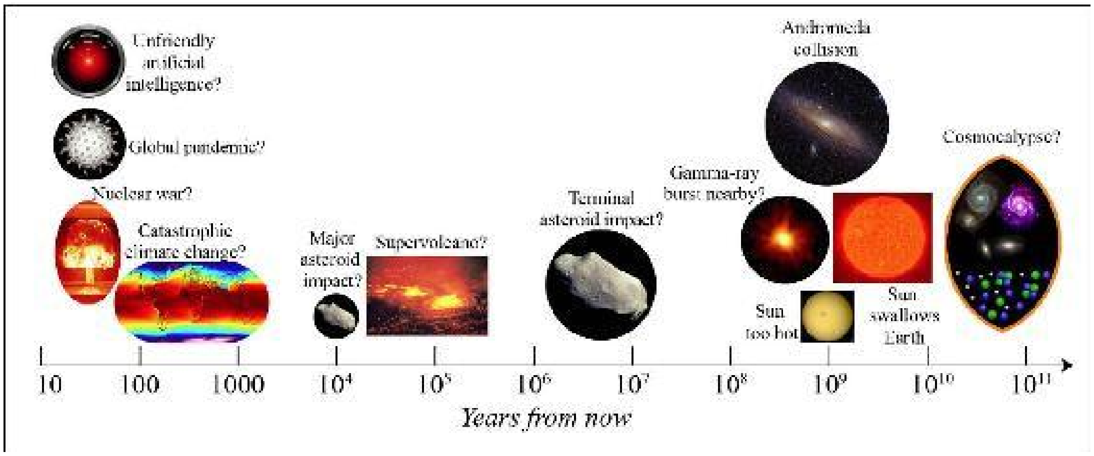
Reversiyon

Durgun totalitarizme boyun eğmeden teknolojinin tehlikelerinden kaçmak cazip olmaz mıydı? Amişlerden esinlenerek ilkel teknolojiye dönerek bunun başarıldığı bir senaryoyu inceleyelim. Omegas, kitabın açılışında olduğu gibi dünyayı ele geçirdikten sonra, basit tarım hayatını romantikleştiren büyük bir küresel propaganda kampanyası başlatıldı.

1500 yıl önce. Dünya nüfusu, teröristleri suçlayan tasarlanmış bir salgın tarafından yaklaşık 100 milyon kişiye düşürüldü. Salgın, bilim veya teknoloji hakkında hiçbir şey bilmeyen kimsenin hayatta kalmamasını sağlamak için gizlice hedef alındı. Prometheus kontrollü robotlar, çok sayıda insanın enfeksiyon tehlikesini ortadan kaldırmak bahanesiyle tüm şehirleri boşalttı ve yerle bir etti. Hayatta kalanlara (aniden elde edilebilecek) geniş arazi parçaları verildi ve yalnızca erken ortaçağ teknolojisini kullanarak sürdürülebilir çiftçilik, balıkçılık ve avcılık uygulamaları konusunda eğitildi. Bu arada, robot orduları sistematik olarak modern teknolojinin tüm izlerini (şehirler, fabrikalar, elektrik hatları ve asfalt yollar dahil) kaldırdı ve bu tür herhangi bir teknolojiyi belgelemeye veya yeniden yaratmaya yönelik tüm insan girişimlerini engelledi. Teknoloji küresel olarak unutulduğunda, robotlar, neredeyse hiçbiri kalmayana kadar diğer robotların sökülmesine yardımcı oldu. En son robotlar, büyük bir termonükleer patlamayla Prometheus'un kendisiyle birlikte kasıtlı olarak buharlaştırıldı. Artık hepsi gittiği için modern teknolojiyi yasaklamaya gerek yoktu. Sonuç olarak, insanlık yapay zeka ya da totalitarizm konusunda endişelenmeden bin yıldan fazla ek süre satın aldı.

Geri dönüş daha önce daha az bir ölçüde gerçekleşti: örneğin, Roma İmparatorluğu sırasında yaygın olarak kullanılan bazı teknolojiler, Rönesans sırasında geri dönüş yapmadan önce yaklaşık bir bin yıl boyunca büyük ölçüde unutulmuştu. Isaac Asimov'un *Yapı temeli* Üçleme, 30.000 yıldan 1.000 yıla kadar bir geri dönüş süresini kısaltmak için "Seldon Planı" etrafında şekilleniyor. Akıllı planlama ile, örneğin tüm tarım bilgisini silerek, tersini yapmak ve bir geri dönüş süresini kısaltmak yerine uzatmak mümkün olabilir. Bununla birlikte, ne yazık ki geri dönüş meraklıları için, bu senaryonun insanlık ya yüksek teknolojiye geçmeden ya da nesli tükenmeden süresiz olarak uzatılabilmesi pek olası değil. Bundan 100 milyon yıl sonra insanların bugünün biyolojik insanlarına benzediğine güvenmek, daha fazlası için bir tür olarak var olmadığımızı düşünürsek, saflık olur.

Şimdiye kadar bu sürenin% 1'inden daha fazla. Dahası, düşük teknolojili insanlık, bir sonraki gezegeni kavuran asteroit çarpması veya Tabiat Ana'nın getirdiği diğer mega felaketler tarafından yok edilmeyi bekleyen savunmasız bir oturan ördek olacaktır. Kesinlikle bir milyar yıl dayanamayız, bundan sonra yavaş yavaş ısınan Güneş, tüm sıvı suyu kaynatacak kadar Dünya'nın sıcaklığını artıracaktır.



Şekil 5.1: Hayatı bildiğimiz haliyle yok edebilecek veya potansiyelini kalıcı olarak kısıtlayabilecek örnekler. Evrenimizin kendisi muhtemelen en az on milyarlarca yıl sürecekken, Güneşimiz yaklaşık bir milyar yıl içinde Dünya'yı kavuracak ve sonra güvenli bir mesafeye götürmezsek onu yutacak ve Gökadamız yaklaşık 3,5 milyar sonra komşusuyla çarpışacak. yıl. Tam olarak ne zaman olduğunu bilmesek de, bundan çok önce asteroitlerin bizi vuracağını ve süper volkanların yıl boyunca güneşsiz kışlara neden olacağını neredeyse kesin olarak tahmin edebiliriz. Teknolojiyi, tüm bu sorunları çözmek veya iklim değişikliği, nükleer savaş, tasarlanmış salgınlar veya ters giden yapay zeka gibi yenilerini yaratmak için kullanabiliriz.

Kendini yok etmek

Gelecekteki teknolojinin neden olabileceği sorunları düşündükten sonra, *eksiklik* bu teknoloji neden olabilir. Bu ruhla, süper zekanın asla yaratılmadığı senaryoları inceleyelim, çünkü insanlık kendini başka yollarla yok ediyor.

Bunu nasıl başarabiliriz? En basit strateji "sadece bekle" dir. Bir sonraki bölümde asteroit etkileri ve kaynayan okyanuslar gibi sorunları nasıl çözebileceğimizi görecek olsak da, bu çözümlerin tümü henüz geliştirmediğimiz teknolojiyi gerektiriyor, bu nedenle teknolojimiz mevcut seviyesinin çok ötesine geçmedikçe Doğa Ana yol gösterecektir. Bir milyar yıl daha geçmeden çok önce soyumuz tükendi. Ünlü ekonomist John Maynard Keynes'in dediği gibi: "Uzun vadede hepimiz ölüyüz."

Ne yazık ki, kolektif aptallık yoluyla kendimizi çok daha erken yok edebileceğimiz yollar da var. Türümüz neden toplu intihar etsin, aynı zamanda *her şeyi öldürmek* neredeyse hiç kimse istemiyorsa? Mevcut zeka ve duygusal olgunluk seviyemizle, biz insanlar yanlış hesaplamalar, yanlış anlamalar ve beceriksizlik konusunda bir ustalığa sahibiz ve sonuç olarak, geçmişimiz geriye dönüp bakıldığında aslında kimsenin istemediği kazalar, savaşlar ve diğer felaketlerle doludur. Ekonomistler ve matematikçiler, insanların eylemlere nasıl teşvik edilebileceğine dair zarif oyun teorisi açıklamaları geliştirdiler.

bu nihayetinde herkes için feci bir sonuca neden olur. [6]

Risk ne kadar büyük olursa o kadar dikkatli olacağımızı düşünebilirsiniz, ancak mevcut teknolojimizin izin verdiği en büyük riskin, yani küresel bir termonükleer savaşın daha yakından incelenmesi güven verici değil. Her türlü şeyin neden olduğu utanç verici derecede uzun bir ramak kala listesi atlatmak için şansa güvenmek zorunda kaldık: bilgisayar arızası, elektrik kesintisi, hatalı istihbarat, navigasyon

hata, bombardıman çarpması, uydu patlaması vb. [7] Aslında, bazı kişilerin kahramanca eylemleri olmasaydı - örneğin, Vasili Arkhipov ve Stanislav Petrov - zaten küresel bir nükleer savaş geçirmiş olabilirdik. Geçmiş performansımıza bakıldığında, eğer mevcut davranışımızı sürdürürsek, yıllık kaza sonucu nükleer savaş olasılığının binde bir kadar düşük olma ihtimalinin çok düşük olduğunu düşünüyorum.

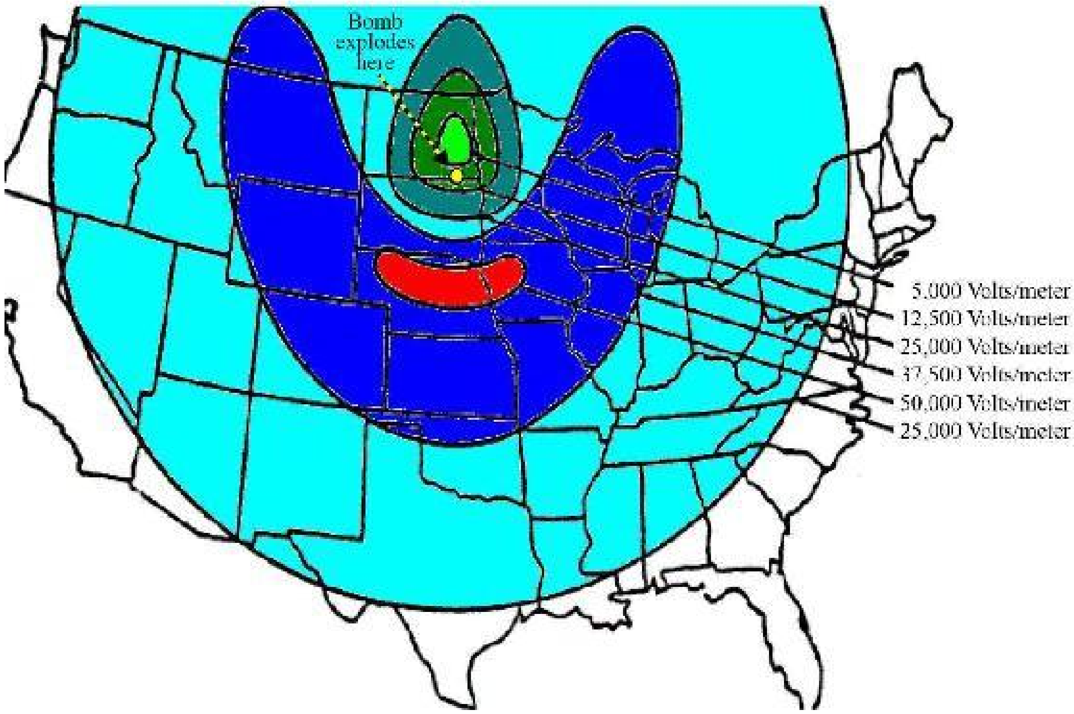
10.000 yıl içinde sahip olma olasılığımız $1 - 0.999$'u aşıyor $^{10000} \approx$ % 99.995.

İnsan pervasızlığımızı tam olarak takdir etmek için, riskleri dikkatlice incelemeden önce bile nükleer kumar oynamaya başladığımızın farkına varmalıyız. İlk olarak, radyasyon riskleri hafife alınmıştı ve uranyum işleme ve nükleer testlerden kaynaklanan radyasyona maruz kalma kurbanlarına 2 milyar doların üzerinde tazminat ödendi.

Yalnızca Birleşik Devletler. [8]

İkincisi, sonunda, hidrojen bombalarının kasıtlı olarak Dünya'nın yüzlerce kilometre yukarısında patlatılmasının, geniş alanlarda elektrik şebekesini ve elektronik cihazları devre dışı bırakabilecek güçlü bir elektromanyetik darbe (EMP) yaratacağı keşfedildi (şekil 5.2), altyapıyı felce uğratmak, yolları engelli araçlarla tıkamak ve nükleer sonrası hayatta kalma koşullarını idealden daha az bırakmak. Örneğin, ABD EMP Komisyonu, "su altyapısının, kısmen yerçekimiyle, ancak çoğunlukla

elektrik, ”ve bu su reddi üç ila dört gün içinde ölüme neden olabilir. [9]



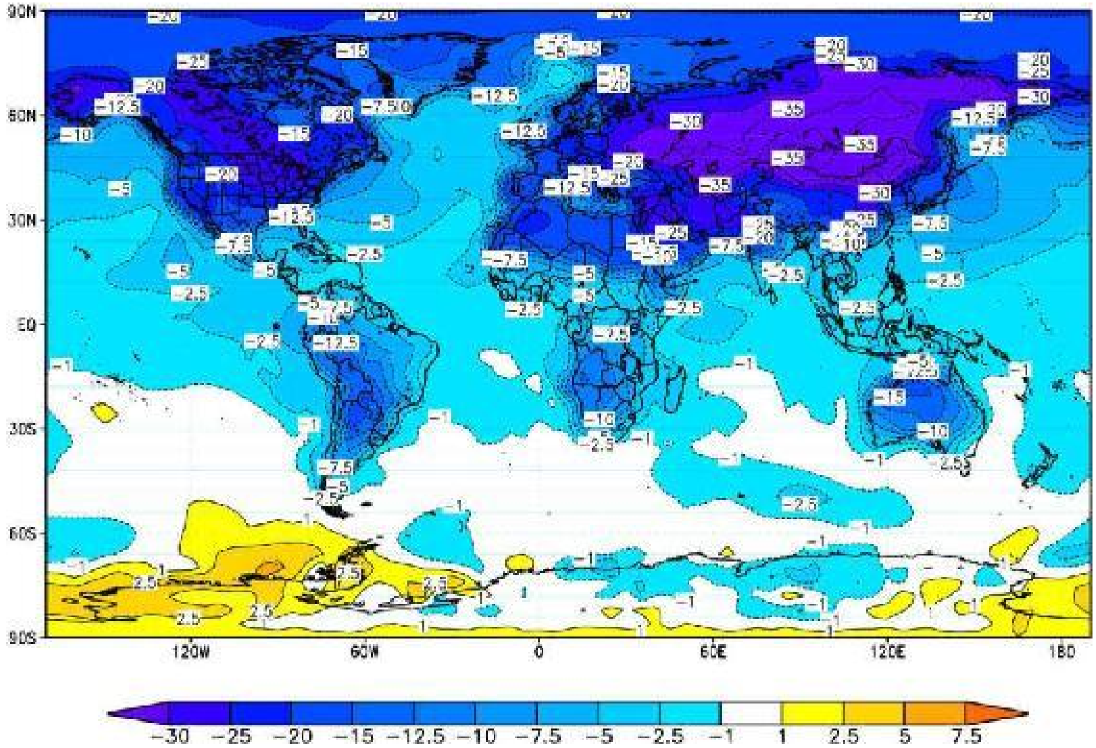
Şekil 5.2: Dünya'nın 400 km yukarısında tek bir hidrojen bombası patlaması, geniş bir alanda elektrik kullanan teknolojiyi bozabilecek güçlü bir elektromanyetik darbeye neden olabilir. Patlama noktasını güneydoğuya kaydırarak, metre başına 37.500 volt'u aşan muz şeklindeki bölge ABD'nin Doğu Kıyısının çoğunu kaplayabilir. ABD Ordusu Raporu AD-A278230'dan (sınıflandırılmamış) renkler eklenmiş olarak yeniden basılmıştır.

Üçüncüsü, nükleer kışın potansiyeli, 63.000 hidrojen bombasını konuşlandırdıktan sonra kırk yıl sonra fark edilmedi - oops! Kimin şehirleri yandığına bakılmaksızın, üst troposfere ulaşan muazzam miktarda duman, tıpkı bir asteroit veya süpervolkanın geçmişte kitlesel bir yok oluşa neden olması gibi, yazları kışa dönüştürmek için yeterli güneş ışığını engelleyerek tüm dünyaya yayılabilir. Alarm 1980'lerde hem ABD hem de Sovyet bilim adamları tarafından çalındığında, bu, Ronald Reagan ve Mikhail Gorbaçov'un kararına katkıda bulundu.

stokları kesmeye başlamak için. [10] Ne yazık ki, daha doğru hesaplamalar daha da kasvetli bir tablo çizdi: şekil 5.3 Amerika Birleşik Devletleri, Avrupa, Rusya ve Çin'in çekirdek tarım bölgelerinin çoğunda (ve Rusya'nın bazı bölgelerinde 35 ° C) ilk ikisi için yaklaşık 20 ° C (36 ° Fahrenheit) kadar soğumayı gösteriyor.

yazlar ve bunun yaklaşık yarısı on yıl sonra bile. *[4] Bu ne anlama geliyor

basit ingilizce? Yaz aylarında neredeyse donmanın gıda üretimimizin çoğunu ortadan kaldıracağı sonucuna varmak için çok fazla çiftçilik deneyimine gerek yok. Dünyanın en büyük binlerce şehri enkaz haline geldikten ve küresel altyapı çöktükten sonra tam olarak ne olacağını tahmin etmek zor, ancak tüm insanların küçük bir kısmı açlığa, hipotermiye veya hastalığa boyun eğmezse, çaresiz kalan silahlı çetelerle başa çıkmak zorunda kalır. Gıda.



Şekil 5.3: Amerika Birleşik Devletleri ve Rusya arasında tam ölçekli bir nükleer savaşın ardından ilk iki yaz boyunca ortalama soğutma (° C cinsinden). Alan Robock'un izniyle çoğaltılmıştır. [11]

Makul bir dünya liderinin istemeyeceği can alıcı noktayı eve götürmek için küresel nükleer savaş hakkında o kadar ayrıntıya girdim ki yine de kazara olabilir. Bu, insan kardeşlerimize asla her şeyi öldürme konusunda güvenemeyeceğimiz anlamına gelir: bunu istemeyen kimse, bunu önlemek için yeterli değildir.

Kıyamet Cihazları

Öyleyse biz insanlar gerçekten öldürücü öldürmeyi başarabilir miyiz? Küresel bir nükleer savaş tüm insanların% 90'ını öldürebilirse bile, çoğu bilim insanı bunun% 100'ünü öldürmeyeceğini ve bu nedenle bizi yok etmeyeceğini tahmin ediyor. Öte yandan, nükleer radyasyon, nükleer EMP ve nükleer kış hikayesi, en büyük tehlikelerin henüz aklımıza bile getirmediğimiz tehlikeler olabileceğini gösteriyor. Sonrasının tüm yönlerini ve nükleer kış, altyapının çöküşü, yükselen mutasyon seviyeleri ve çaresiz silahlı orduların yeni salgınlar, ekosistem çökmesi ve henüz hayal etmediğimiz etkiler gibi diğer sorunlarla nasıl etkileşime girebileceğini öngörmek inanılmaz derecede zor. Bu nedenle kişisel değerlendirmem, yarın insan neslinin tükenmesini tetikleyecek bir nükleer savaş olasılığı büyük olmasa da, bunun da sıfır olduğu sonucuna güvenle varamayız.

Bugünün nükleer silahlarını kasıtlı bir kıyamet günü cihazına yükseltirsek, omnicide oranları artar. 1960 yılında RAND stratejisti Herman Kahn tarafından tanıtıldı ve Stanley Kubrick'in filminde popüler oldu *Dr. Strangelove,* bir kıyamet günü cihazı, karşılıklı olarak garantili yıkım paradigmasını nihai sonucuna götürür. Bu mükemmel bir caydırıcı: tüm insanlığı öldürerek herhangi bir düşman saldırısına otomatik olarak misilleme yapan bir makine.

Kıyamet günü cihazı için bir aday, sözde büyük bir yeraltı önbelleğidir. *tuzlu nükleer silahlar* tercihen devasa miktarda kobaltla çevrili devasa hidrojen bombaları. Fizikçi Leo Szilard, 1950'de bunun Dünya'daki herkesi öldürebileceğini savundu: Hidrojen bombası patlamaları kobalt radyoaktif hale getirecek ve stratosfere üfleyecekti ve beş yıllık yarı ömrü tüm Dünya'ya yerleşecek kadar uzun. (özellikle ikiz kıyamet cihazları zıt yarıkürelere yerleştirilmişse), ancak ölümcül radyasyon yoğunluğuna neden olacak kadar kısa. Basında çıkan haberler, kobalt bombalarının artık ilk kez üretildiğini gösteriyor. Stratosferde uzun ömürlü aerosolleri maksimize ederek nükleer kış oluşumu için optimize edilmiş bombalar eklenerek omnisidal fırsatlar artırılabilir. Bir kıyamet günü cihazının en önemli satış noktası, geleneksel bir nükleer caydırıcıdan çok daha ucuz olmasıdır: çünkü bombaların fırlatılmasına gerek yoktur,

Diğer bir olasılık, biyolojik bir kıyamet günü cihazının gelecekteki keşfidir:

tüm insanları öldüren özel tasarlanmış bakteri veya virüs. Eğer bulaşıcılığı yeterince yüksekse ve kuluçka süresi yeterince uzun olsaydı, esasen herkes onu varlığını fark etmeden ve karşı önlemler almadan önce yakalayabilirdi. Herkesi öldüremese bile böyle bir biyolojik silah inşa etmek için askeri bir argüman var: en etkili kıyamet günü cihazı, düşmanı caydırma şansını en üst düzeye çıkarmak için nükleer, biyolojik ve diğer silahları birleştiren cihazdır.

AI Silahları

Omnicide giden üçüncü bir teknolojik yol, nispeten aptal AI silahları içerebilir. Bir süper gücün 3. bölümden milyarlarca yaban arısı büyüklüğünde saldırı dronu ürettiğini ve bunları, tıpkı günümüz süpermarket ürünlerinin çoğu gibi uzaktan bir radyo frekansı kimlik etiketi ile tanımlanan kendi vatandaşları ve müttefikleri dışındaki herkesi öldürmek için kullandığını varsayalım. Bu etiketler, totalitarizm bölümünde olduğu gibi tüm vatandaşlara bileklik takmak veya transdermal implant olarak dağıtılabilir. Bu muhtemelen muhalif bir süper gücü benzer bir şey inşa etmeye teşvik edecektir. Savaş kazara patlak verdiğinde, tüm insanlar, hatta bağlı olmayan uzak kabileler bile öldürülecekti, çünkü hiç kimse bu iki tür kimlik kartını takmayacaktır. Bunu nükleer ve biyolojik bir kıyamet günü cihazı ile birleştirmek, başarılı omnicide şansını daha da artıracaktır.

Ne Yapar *Sen* İstemek?

Bu bölüme, mevcut AGI yarışının nereye varmasını istediğinizi düşünerek başladınız. Artık geniş bir senaryo yelpazesini birlikte araştırdığımıza göre, hangileri size hitap ediyor ve hangilerinden kaçınmak için çok uğraşmamız gerektiğini düşünüyorsunuz? Net bir favoriniz var mı? Lütfen bana ve diğer okuyucu arkadaşlarımdan haberdar edin http://AgeOfAi.org ve tartışmaya katılın!

Açıkça ele aldığımız senaryolar tam bir liste olarak görülmemelidir ve çoğu ayrıntıda zayıftır, ancak kapsayıcı olmak için çok çalıştım, yüksek teknolojiden düşük teknolojiye ve teknolojisiz ve literatürde ifade edilen tüm merkezi umutları ve korkuları açıklamak.

Bu kitabı yazmanın en eğlenceli kısımlarından biri, arkadaşlarımın ve meslektaşlarımın bu senaryolar hakkında ne düşündüğünü duymak oldu ve hiçbir şekilde fikir birliği olmadığını öğrenmekten keyif aldım. Herkesin hemfikir olduğu tek şey, seçimlerin başlangıçta göründüğünden daha ince olmasıdır. Herhangi bir senaryoyu beğenen insanlar, aynı anda bazı yönlerini rahatsız edici buluyorlar. Bana göre bu, biz insanların gelecekteki hedeflerimiz hakkındaki bu sohbeti sürdürmemiz ve derinleştirmemiz gerektiği anlamına geliyor, böylece hangi yöne gideceğimizi biliyoruz. Evrenimizdeki yaşamın gelecekteki potansiyeli hayranlık uyandıracak kadar büyüktür, bu yüzden nereye gitmek istediğimiz konusunda hiçbir fikri olmayan dümensiz bir gemi gibi sürüklenerek onu israf etmeyelim!

Bu gelecekteki potansiyel ne kadar büyük? Teknolojimiz ne kadar gelişirse gelişsin, Life 3.0'ın gelişmesi ve kozmosumuza yayılması, fizik yasalarıyla sınırlı olacaktır - önümüzdeki milyarlarca yıl boyunca bu nihai sınırlar nelerdir? Evrenimiz şu anda dünya dışı yaşamla dolu mu yoksa yalnız mıyız? Farklı genişleyen kozmik medeniyetler bir araya gelirse ne olur? Bir sonraki bölümde bu büyüleyici soruları ele alacağız.

ALT ÇİZGİ:

- AGI'ye yönelik mevcut yarış, önümüzdeki bin yıl için şaşırtıcı derecede geniş bir sonraki senaryo yelpazesiyle sona erebilir.

- Süper zeka, zorlandığı için (köleleştirilmiş tanrı senaryosu) ya da isteyen “dost canlısı yapay zeka” olduğu için (özgürlükçü-ütopya, koruyucu-tanrı, yardımsever-diktatör ve zookeeper senaryoları) insanlarla barış içinde bir arada var olabilir.

- Süper zeka, bir yapay zeka (bekçi senaryosu) veya insanlar tarafından (1984 senaryosu), kasıtlı olarak teknolojiyi unutarak (tersine çevirme senaryosu) veya onu inşa etmek için teşviklerin olmaması (eşitlikçi-ütopya senaryosu) tarafından önlenebilir.

- İnsanlığın soyu tükenebilir ve yapay zekalar (fatih ve alt senaryolar) veya hiçbir şey (kendi kendini yok etme senaryosu) ile değiştirilebilir.

- Bu senaryolardan hangisinin istenirse, hangisinin arzu edildiği konusunda kesinlikle bir fikir birliği yoktur ve tümü sakıncalı unsurlar içerir. Bu, gelecekteki hedeflerimiz etrafında sohbeti sürdürmeyi ve derinleştirmeyi daha da önemli hale getirir, böylece istemeden talihsiz bir yöne kaymaz veya yön vermeyiz.

---

* 1 Bu fikir, "Bir şey başkalarıyla paylaşılarak eksiltilmezse, sadece sahip olunursa ve paylaşılmazsa hak sahibi olunmaz" yazan Saint Augustine'e kadar uzanıyor.

* 2 Bu fikir bana ilk olarak arkadaşım ve meslektaşım Anthony Aguirre tarafından önerildi.

* 3 Ünlü kozmolog Fred Hoyle, İngiliz TV dizisinde farklı bir bükülme ile ilgili bir senaryoyu araştırdı. *Andromeda için.*

* 4 Atmosfere karbon enjekte etmek iki tür iklim değişikliğine neden olabilir: karbondioksitten ısınma veya duman ve kurumdan soğutma. Bilimsel kanıt olmadan ara sıra reddedilen sadece ilk tür değil: Bazen bana nükleer kışın çürütüldüğü ve neredeyse imkansız olduğu söylendi. Her zaman, böylesine güçlü iddialarda bulunan hakemli bir bilimsel makaleye atıfta bulunarak yanıt veriyorum ve şimdiye kadar hiçbiri yok gibi görünüyor. Özellikle ne kadar dumanın üretildiği ve ne kadar yükseldiği ile ilgili daha fazla araştırmayı gerektirecek büyük belirsizlikler olsa da, benim bilimsel görüşüme göre nükleer kış riskini göz ardı etmek için geçerli bir dayanak yok.

Bölüm 6

Kozmik Bağışımız: Sonraki Milyar Yıllar ve Ötesi

Spekülasyonumuz bir süper medeniyetle, tüm güneş sistemi yaşamının sentezi, kendini sürekli geliştiren ve genişleten, güneşten dışarıya doğru yayılan, hayatı olmayanları akla dönüştüren bir süper medeniyetle bitiyor.

Hans Moravec, *Mind Children*

Bana göre, şimdiye kadarki en ilham verici bilimsel keşif, yaşamın gelecekteki potansiyelini önemli ölçüde hafife almış olmamızdır. Hayallerimizin ve özlemlerimizin hastalık, yoksulluk ve kafa karışıklığıyla gölgelenmiş yüzyıllık yaşam süreleriyle sınırlı olması gerekmez. Aksine, teknolojinin yardımıyla, hayatın milyarlarca yıl boyunca gelişme potansiyeli var, sadece burada Güneş Sistemimizde değil, aynı zamanda atalarımızın hayal ettiğinden çok daha büyük ve ilham verici bir kozmosta. Gökyüzü bile sınır değil.

Bu, çağlar boyunca sınırları zorlayarak ilham alan bir tür için heyecan verici bir haber. Olimpiyat oyunları güç, hız, çeviklik ve dayanıklılığın sınırlarını zorlayarak kutlar. Bilim, bilgi ve anlayışın sınırlarını zorlamayı kutluyor. Edebiyat ve sanat, güzel veya yaşamı zenginleştiren deneyimler yaratmanın sınırlarını zorlayarak kutluyor. Birçok kişi, kuruluş ve ülke artan kaynakları, bölgeleri ve uzun ömürlülüğü kutluyor. Sınırlara olan insan takıntımız göz önüne alındığında, tüm zamanların en çok satan telif hakkıyla korunan kitabının *Guinness Rekorlar Kitabı.*

Öyleyse, eski algılanan yaşam sınırlarımız teknoloji tarafından paramparça edilebilirse,

nihai limitler? Kozmosumuzun ne kadarı canlanabilir? Hayat ne kadar uzağa ulaşabilir ve ne kadar sürebilir? Hayat ne kadar maddeden yararlanabilir ve ne kadar enerji, bilgi ve hesaplama çıkarabilir? Bu nihai sınırlar bizim anlayışımızla değil, fizik yasalarıyla belirlenir. Bu, ironik bir şekilde, hayatın uzun vadeli geleceğini analiz etmeyi kısa vadeli geleceğe göre bazı yönlerden daha kolay hale getiriyor.

13,8 milyar yıllık kozmik tarihimiz bir haftaya sıkıştırılsaydı,
Son iki bölümün 10.000 yıllık draması yarım saniyeden daha kısa sürede biter. Bu, bir istihbarat patlamasının ortaya çıkıp çıkmayacağını ve nasıl ortaya çıkacağını ve hemen sonrasının nasıl olacağını kestirememememize rağmen, tüm bu kargaşanın kozmik tarihte sadece ayrıntıları yaşamın nihai sınırlarını etkilemeyen kısa bir flaş olduğu anlamına gelir. Patlama sonrası yaşam, günümüz insanlarının sınırları zorlamak için takıntılıysa, o zaman teknolojiyi geliştirecektir. *ulaşmak* bu sınırlar

- çünkü olabilir. Bu bölümde, bu sınırların ne olduğunu keşfedeceğiz, böylece yaşamın uzun vadeli geleceğinin nasıl olabileceğine bir göz atacağız. Bu sınırlar, mevcut fizik anlayışımıza dayandığından, olasılıkların alt sınırı olarak görülmeleri gerekir: gelecekteki bilimsel keşifler daha da iyisini yapma fırsatları sunabilir.

Ama gelecekteki yaşamın bu kadar hırslı olacağını gerçekten biliyor muyuz? Hayır, yapmayız: Belki de bir eroin bağımlısı ya da sırf bitmeyen tekrarları izleyen bir kanepe patates kadar kayıtsız hale geleceğiz. *Kardashians ile tutmak.* Bununla birlikte, hırsın ileri yaşamın oldukça genel bir özelliği olduğundan şüphelenmek için nedenler var. Neredeyse maksimize etmeye çalıştığı şey ne olursa olsun, zeka, uzun ömürlülük, bilgi veya ilginç deneyimler, kaynaklara ihtiyacı olacaktır. Bu nedenle, sahip olduğu kaynaklardan en iyi şekilde yararlanmak için teknolojisini nihai sınırlara kadar zorlama teşviki vardır. Bundan sonra, daha da gelişmenin tek yolu, kozmosun her zamankinden daha geniş bölgelerine genişleyerek daha fazla kaynak elde etmektir.

Ayrıca yaşam, kozmosumuzun birçok yerinde bağımsız olarak ortaya çıkabilir. Bu durumda, hırslı medeniyetler kozmik olarak önemsiz hale gelir ve kozmik bağışın giderek daha büyük kısımları nihayetinde en hırslı yaşam formları tarafından ele geçirilir. Bu nedenle doğal seçilim, kozmik ölçekte oynar ve bir süre sonra var olan neredeyse tüm yaşam hırslı yaşam olacaktır. Özetle, kozmosumuzun nihayetinde ne ölçüde canlanabileceği ile ilgileniyorsak, fizik yasalarının dayattığı hırsın sınırlarını incelemeliyiz. Bunu yapalım! Önce Güneş Sistemimizde sahip olduğumuz kaynaklarla (madde, enerji vb.) Neler yapılabileceğinin sınırlarını inceleyelim, sonra kozmik keşif ve yerleşim yoluyla nasıl daha fazla kaynak elde edebileceğimize dönelim.

Kaynaklarınızdan En İyi Şekilde Yararlanmak

Bugünün süpermarketleri ve borsaları "kaynaklar" olarak adlandırabileceğimiz on binlerce ürün satarken, teknolojik sınıra ulaşan gelecek yaşam temelde bir temel kaynağa ihtiyaç duyar: sözde *baryonik madde*

atomlardan veya bileşenlerinden (kuarklar ve elektronlar) oluşan herhangi bir şey anlamına gelir. Bu madde ne şekilde olursa olsun, ileri teknoloji onu enerji santralleri, bilgisayarlar ve ileri yaşam formları dahil olmak üzere istenen maddelere veya nesnelere yeniden düzenleyebilir. Öyleyse, ileri yaşama güç veren enerjinin sınırlarını ve onun düşünmesini sağlayan bilgi işlemeyi inceleyerek başlayalım.

Dyson Küreleri Oluşturma

Hayatın geleceği söz konusu olduğunda, en umutlu vizyonerlerden biri Freeman Dyson'dur. Son yirmi yıldır onu tanımanın şeref ve zevkini yaşadım, ama onunla ilk tanıştığımda gergin hissettim. Princeton'daki İleri Araştırma Enstitüsü'nün yemekhanesinde arkadaşlarımla yemek yiyen küçük bir doktora sonrası öğrenciydim ve birdenbire, Einstein ve Gödel ile takılan dünyaca ünlü fizikçi gelip kendini tanıttı ve sordu: bize katılabilir! Bununla birlikte, gençlerle öğle yemeği yemeyi havasız eski profesörlere tercih ettiğini açıklayarak beni çabucak rahatlattı. Ben bu kelimeleri yazarken doksan üç yaşında olmasına rağmen, Freeman ruh olarak tanıdığım çoğu insandan daha genç ve gözlerindeki yaramaz çocuksu parıltı formaliteleri daha az umursayamayacağını ortaya koyuyor. akademik hiyerarşiler veya geleneksel bilgelik. Fikir ne kadar cesursa, o kadar heyecanlanır.

Enerji kullanımı hakkında konuştuğumuzda, biz insanların ne kadar hırslı olduğumuzla alay etti ve Sahra çölünün% 0,5'inden daha küçük bir alana çarpan güneş ışığını toplayarak mevcut tüm küresel enerji ihtiyacımızı karşılayabileceğimize dikkat çekti. Ama neden orada duralım? Neden Dünya'ya çarpan tüm güneş ışığını yakalamayı bırakıp çoğunun boşluğa savurgan bir şekilde ışınlanmasına izin verelim? Neden basitçe söylemek değil

herşey Güneşin enerji çıkışı yaşam için kullanmak için?

Olaf Stapledon'un 1937 bilim kurgu klasiğinden esinlenilmiştir *Star Maker,* ana yıldızlarının etrafında dönen yapay dünyaların halkaları ile Freeman Dyson, bir

1960 yılında bir *Dyson küresi.* [1] Freeman'ın fikri, Jüpiter'i Güneş'i çevreleyen küresel bir kabuk şeklinde bir biyosferde yeniden düzenlemek ve torunlarımızın gelişip 100 milyarın tadını çıkarmaktı.

insanlığın bugün kullandığından kat kat fazla biyokütle ve trilyon kat daha fazla enerji. [2]

Bunun doğal bir sonraki adım olduğunu savundu: "Sanayileşme aşamasına girdikten sonraki birkaç bin yıl içinde, herhangi bir akıllı türün, ana yıldızını tamamen çevreleyen yapay bir biyosferde bulunması gerektiğini beklemek gerekir." Bir Dyson küresinin içinde yaşıyor olsaydınız, gece olmazdı: Güneş'i her zaman tam tepenizde görürdünüz ve tüm gökyüzünde güneş ışığının, tıpkı yapabildiğiniz gibi biyosferin geri kalanından yansıdığını görürsünüz. günümüzde gün boyunca Ay'dan yansıyan güneş ışığı görüyor. Yıldızları görmek istiyorsanız, sadece "yukarı" çıkıp Dyson küresinin dışından evrene bakarsınız.

Kısmi bir Dyson küresi oluşturmanın düşük teknolojili bir yolu, Güneş'in etrafındaki dairesel yörüngeye bir habitat halkası yerleştirmektir. Güneşi tamamen çevrelemek için, çarpışmalardan kaçınmak için farklı eksenler etrafında biraz farklı mesafelerde dönen halkalar ekleyebilirsiniz. Bu hızlı hareket eden halkaların birbirine bağlanamaması, ulaşım ve iletişimi karmaşık hale getirme sıkıntısından kaçınmak için, Güneş'in içe doğru yerçekimi kuvvetinin Güneş'in radyasyonundan gelen dışa doğru basınçla dengelendiği yekpare bir sabit Dyson küresi inşa edilebilir. Robert L. Forward ve Colin McInnes'in öncülük ettiği bir fikir. Küre, kademeli olarak daha fazla “statit” eklenerek inşa edilebilir: Güneş'in yerçekimini merkezkaç kuvvetlerden ziyade radyasyon basıncı ile etkisiz hale getiren sabit uydular. Bu kuvvetlerin her ikisi de Güneş'e olan uzaklığın karesiyle düşer, Bu, Güneş'ten bir mesafede dengelenebilirse, başka herhangi bir mesafede de rahatça dengelenecekleri anlamına gelir ve Güneş Sistemimizdeki herhangi bir yere park etme özgürlüğü sağlar. Statitlerin yalnızca ağırlıkta olan son derece hafif tabakalar olması gerekir

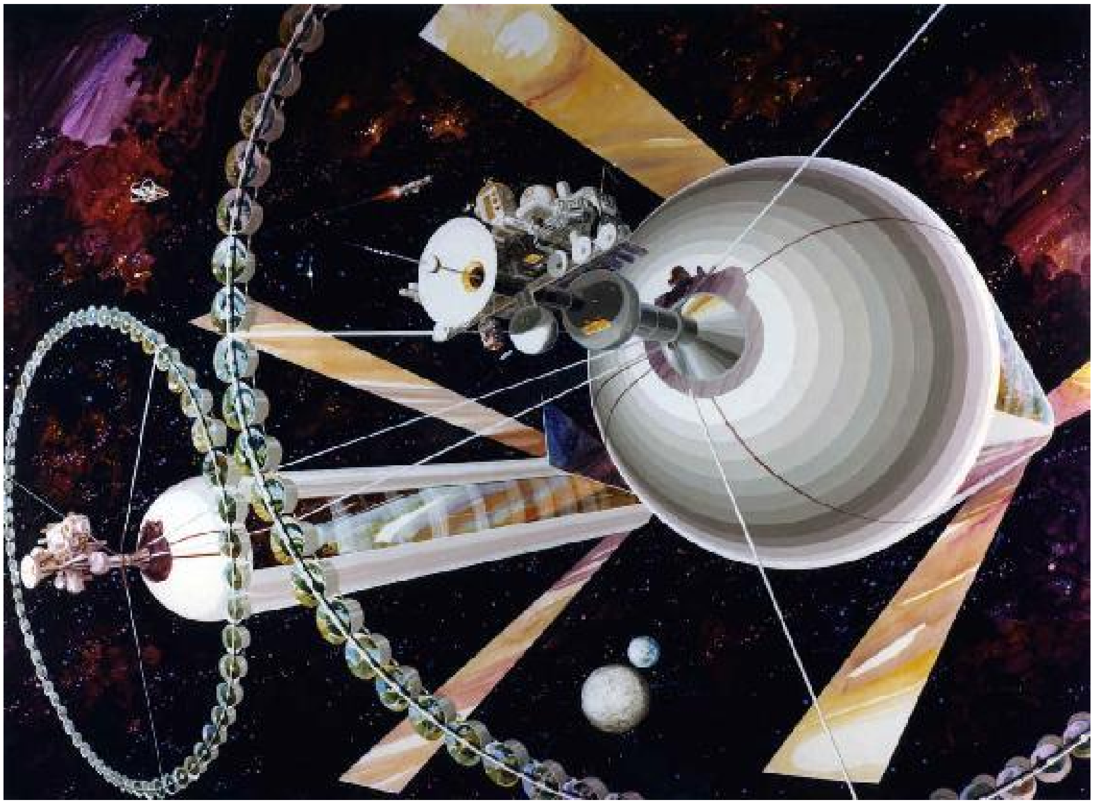
Metrekare başına 0.77 gram, bu kağıttan yaklaşık 100 kat daha az, ancak bunun bir gösterici olması pek olası değil. Örneğin, bir grafen tabakası (tavuk teline benzeyen altıgen bir desende tek bir karbon atomu tabakası) bu sınırın bin katı ağırlığındadır. Dyson küresi güneş ışığının çoğunu absorbe etmek yerine yansıtacak şekilde inşa edilirse, etrafından sıçrayan toplam ışık yoğunluğu önemli ölçüde artacak ve bu da radyasyon basıncını ve kürede desteklenebilecek kütle miktarını daha da artıracaktır. Diğer pek çok yıldız, Güneşimizden bin kat ve hatta milyon kat daha fazla parlaklığa sahiptir ve bu nedenle buna uygun olarak daha ağır sabit Dyson kürelerini destekleyebilir.

Burada Güneş Sistemimizde çok daha ağır sert bir Dyson küresi isteniyorsa, Güneş'in yerçekimine direnmek, sıvılaşmadan veya dünyanın en yüksek gökdelenlerinin tabanındakilerden on binlerce kat daha fazla basınçlara dayanabilecek ultra güçlü malzemeler gerektirecektir. burkulma. Uzun ömürlü olması için, bir Dyson küresinin dinamik ve akıllı olması, rahatsızlıklara yanıt olarak konumunu ve şeklini sürekli olarak ince ayarlaması ve bazen rahatsız edici asteroitlerin ve kuyruklu yıldızların olaysız geçmesine izin vermek için büyük delikler açması gerekir. Alternatif olarak, bu tür sistem davetsiz misafirlerini idare etmek, isteğe bağlı olarak onları parçalara ayırmak ve maddelerini daha iyi kullanmak için bir algılama ve saptırma sistemi kullanılabilir.

Bugünün insanları için, bir Dyson küresindeki ya da içindeki yaşam en iyi ihtimalle kafa karıştırıcı ve en kötü ihtimalle imkansız olacaktır, ancak bunun gelecekteki biyolojik ya da biyolojik olmayan yaşam formlarının orada gelişmesini durdurması gerekmez. Yörüngedeki varyant aslında hiç yerçekimi sunmazdı ve eğer durağan türden etrafta dolaşırsanız, düşmeden sadece dışarıdan (Güneş'e bakmadan) yürüyebilirsiniz,

alışık olduğunuzdan yaklaşık on bin kat daha zayıf yerçekimi ile. Sizi Güneş'ten gelen tehlikeli parçacıklardan koruyan bir manyetik alanınız olmayacak (bir tane oluşturmadıysanız). İşin güzel yanı, Dünya'nın şu anki yörüngesi büyüklüğünde bir Dyson küresinin bize üzerinde yaşamak için yaklaşık 500 milyon kat daha fazla yüzey alanı vermesidir.

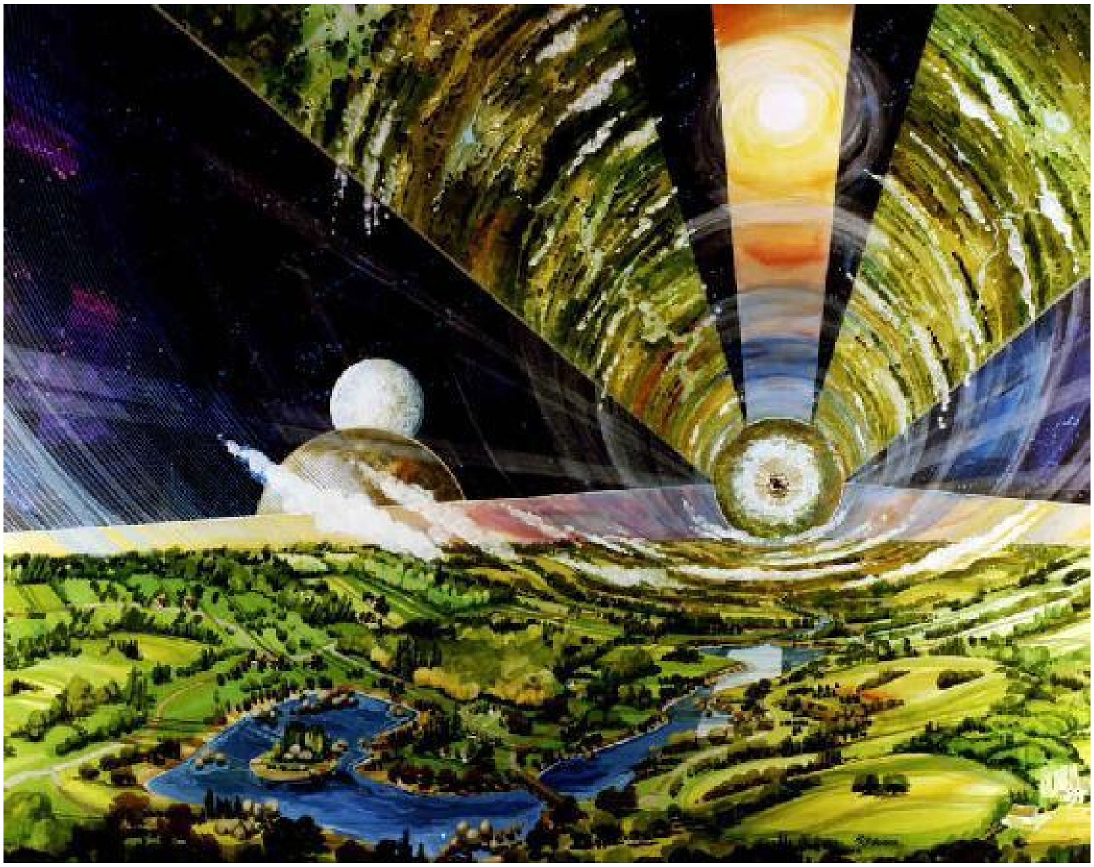
Daha fazla Dünya benzeri insan yaşam alanı isteniyorsa, iyi haber, bunların bir Dyson küresinden çok daha kolay inşa edilmesidir. Örneğin, rakamlar 6.1 ve 6.2 Yapay yerçekimini, kozmik ışın korumasını, yirmi dört saatlik gündüz-gece döngüsünü ve Dünya benzeri atmosfer ve ekosistemleri destekleyen Amerikalı fizikçi Gerard K. O'Neill'in öncülüğünü yaptığı silindirik bir habitat tasarımını gösterin. Bu tür habitatlar, bir Dyson küresinin içinde serbestçe yörüngede dolaşabilir veya değiştirilmiş varyantlar onun dışına eklenebilir.



Şekil 6.1: Bir çift ters dönen O'Neill silindiri, Güneş'in etrafında her zaman doğrudan onu gösterecek şekilde yörüngede dönerlerse, Dünya benzeri rahat yaşam alanları sağlayabilir. Dönüşlerinden kaynaklanan merkezkaç kuvveti yapay yerçekimi sağlar ve üç katlanabilir ayna, 24 saatlik bir gündüz-gece döngüsünde içeriye güneş ışığı gönderir. Bir halka şeklinde düzenlenmiş daha küçük habitatlar, tarım için uzmanlaşmıştır. Resim Rick Guidice / NASA'nın izniyle.

Daha İyi Santraller Kurmak

Dyson küreleri günümüzün mühendislik standartlarına göre enerji açısından verimli olsalar da, fizik kanunlarının koyduğu sınırları zorlayacak kadar yakın bir noktaya gelmiyorlar. Einstein bize öğretti ki kütleyi enerjiye% 100 verimlilikle dönüştürebilirsek * [1] sonra bir miktar kütle *m* bize bir miktar enerji verirdi *E* ünlü tarafından verildi formül $E = mc^2$, nerede *c* ışık hızıdır. Bu, o zamandan beri *c* çok büyükse, az miktarda kütle muazzam miktarda enerji üretebilir. Bol miktarda antimadde kaynağımız olsaydı (ki bizde yok), o zaman% 100 verimli bir elektrik santrali yapmak kolay olurdu: normal suya bir çay kaşığı anti-su dökmek, 200.000 ton TNT'ye eşdeğer enerjiyi açığa çıkarırdı. , tipik bir hidrojen bombasının verimi - yaklaşık yedi dakika boyunca dünyanın tüm enerji ihtiyacını karşılamaya yetecek kadar.



Şekil 6.2: Önceki şekilden O'Neill silindirlerinden birinin iç görünümü. Çapı 6.4 kilometre ise ve her 2 dakikada bir dönüyorsa, yüzeydeki insanlar Dünya'dakiyle aynı görünür yerçekimini yaşayacaktır. Güneş arkanızda, ancak geceleri katlanan silindirin dışındaki bir ayna nedeniyle yukarıda beliriyor. Hava geçirmez pencereler, atmosferin silindirden kaçmasını engeller. Resim Rick Guidice / NASA'nın izniyle.

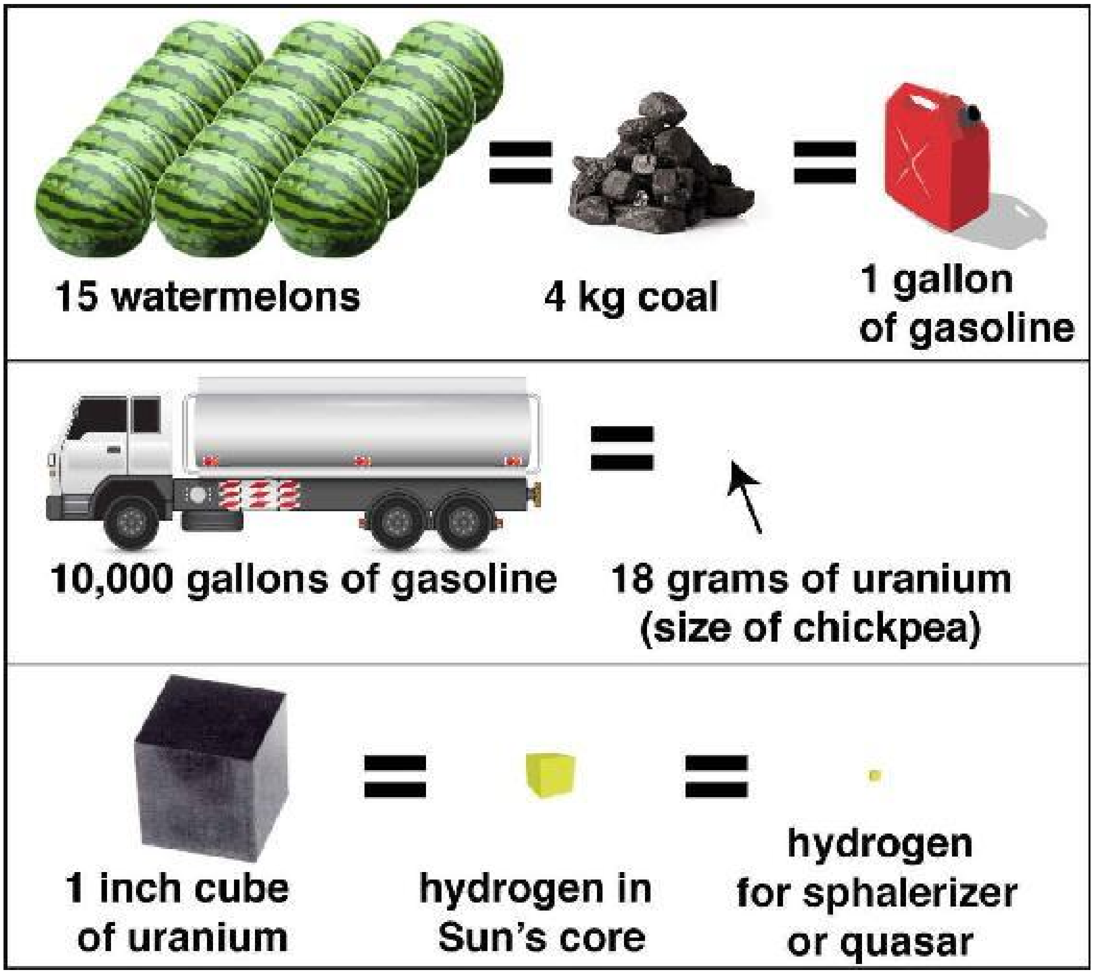
Buna karşılık, günümüzde en yaygın enerji üretme yöntemlerimiz, aşağıda özetlendiği gibi, ne yazık ki verimsizdir. tablo 6.1 ve şekil 6.3 . Bir çikolatayı sindirmek, yalnızca on trilyonda birini serbest bırakması açısından yalnızca% 0,00000001 etkilidir.

enerjinin mc^2 içerdiği. Mideniz% 0,001 bile verimli olsaydı, hayatınızın geri kalanında yalnızca tek bir öğün yemeniz gerekirdi. Yemek yemeye kıyasla, kömür ve benzinin yakılması sırasıyla sadece 3 ve 5 kat daha verimli. Bugünün nükleer reaktörleri uranyum atomlarını fisyon yoluyla bölerek önemli ölçüde daha iyi sonuç verirken, yine de enerjilerinin% 0,08'inden fazlasını çıkarmada başarısız oluyor. Güneş'in çekirdeğindeki nükleer reaktör çok daha verimli

bizim inşa ettiklerimizden daha fazla, enerjinin% 0.7'sini hidrojenden helyuma kaynaştırarak çıkarıyor. Bununla birlikte, Güneş'i mükemmel bir Dyson küresi içine alsak bile, Güneş kütlesinin yaklaşık% 0,08'inden fazlasını kullanabileceğimiz enerjiye asla dönüştüremeyiz çünkü Güneş hidrojen yakıtının yaklaşık onda birini tükettiğinde, normal bir yıldız olarak ömrünü sona erdirecek, kırmızı bir deve dönüşecek ve ölmeye başlayacaktır. Diğer yıldızlar için de işler daha iyi hale gelmiyor: Ana yaşamları boyunca tükettikleri hidrojenin oranı, çok küçük yıldızlar için yaklaşık% 4'ten en büyük yıldızlar için yaklaşık% 12'ye kadar değişiyor. Elimizdeki tüm hidrojenin% 100'ünü birleştirmemize izin verecek yapay bir füzyon reaktörünü mükemmelleştirirsek, yine de füzyon işleminin utanç verici derecede düşük% 0.7'lik verimliliğine takılıp kalırız. Nasıl daha iyi yapabiliriz?

Yöntem	**Verimlilik**
Sindiren şeker çubuğu	% 0.00000001
Yanan kömür	% 0.00000003
Yanan benzin	% 0.00000005
Uranyum-235'in bölünmesi	% 0,08
Güneş ölünceye kadar Dyson küresini kullanmak	% 0,08
Hidrojenin helyuma füzyonu Kara delik motoru	% 0.7
döndürmek	% 29
Quasar Sphalerizer çevresinde Dyson	% 42
küre	% 50 mi?
Kara delik buharlaşması	% 90

Tablo 6.1: Kütlenin teorik sınıra göre kullanılabilir enerjiye dönüştürülmesinin etkinliği $E = mc^2$. Metinde açıklandığı gibi, kara delikleri besleyerek% 90 verim elde etmek ve buharlaşmalarını beklemek ne yazık ki yararlı olamayacak kadar yavaştır ve sürecin hızlandırılması verimliliği önemli ölçüde düşürür.



Şekil 6.3: İleri teknoloji, maddeden onu yiyerek veya yakarak elde ettiğimizden çok daha fazla enerji çıkarabilir ve hatta nükleer füzyon, fizik kanunlarının belirlediği sınırlardan 140 kat daha az enerji çıkarabilir. Sfaleronları, kuasarları veya buharlaşan kara delikleri kullanan enerji santralleri çok daha iyisini yapabilir.

Buharlaşan Kara Delikler

Kitabında *Zamanın Kısa Tarihi,* Stephen Hawking bir kara delik önerdi

enerji santrali. * 2 Kara deliklerin uzun zamandır ışığın bile kaçamayacağı tuzaklar olduğuna inanılırsa kulağa paradoksal gelebilir. Bununla birlikte, Hawking, kuantum yerçekimi etkilerinin bir kara deliğin sıcak bir nesne gibi hareket etmesini sağladığını (ne kadar küçükse, o kadar sıcaksa), şu anda bilinen ısı radyasyonu yayan meşhur hesaplamıştır. *Hawking radyasyonu.* Bu, kara deliğin yavaş yavaş enerji kaybettiği ve buharlaştığı anlamına gelir. Başka bir deyişle, kara deliğe attığınız her ne olursa olsun, sonunda ısı radyasyonu olarak geri gelecektir, bu nedenle kara delik tamamen buharlaştığında, siz

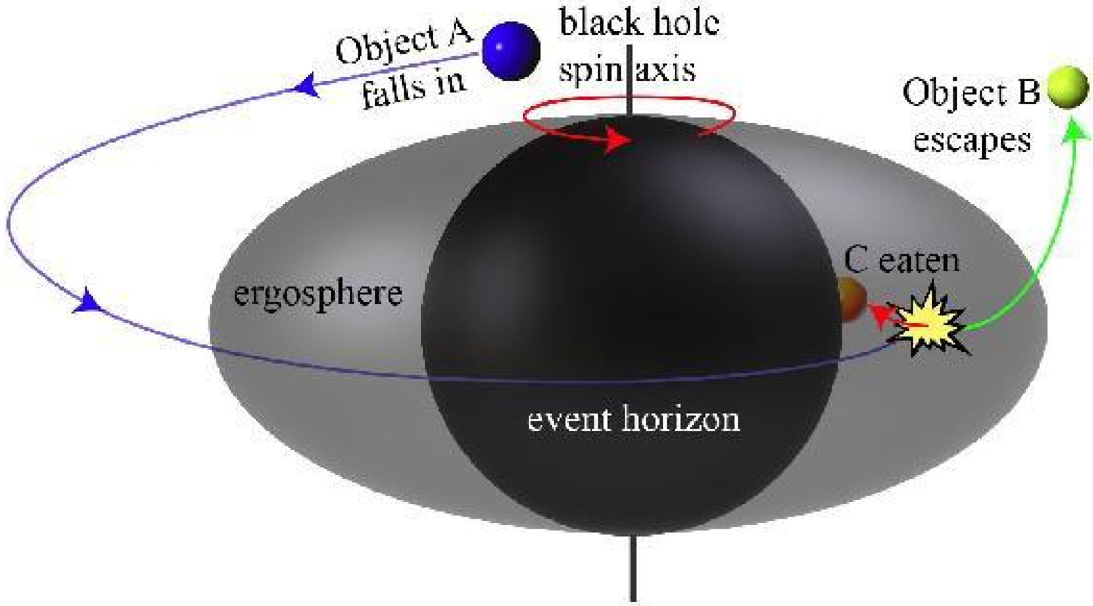
yaklaşık% 100 verimlilikle maddenizi radyasyona dönüştürdü. * 3

Kara delik buharlaşmasını güç kaynağı olarak kullanmanın bir problemi, kara delik boyut olarak bir atomdan çok daha küçük olmadığı sürece, Evrenimizin şu anki yaşından daha uzun süren ve bir mumdan daha az enerji yayan, dayanılmaz derecede yavaş bir süreçtir. Üretilen güç, deliğin karesi boyutuna göre azalır ve fizikçiler Louis Crane ve Shawn Westmoreland bu nedenle bir karadelikten yaklaşık bin kat daha küçük bir kara delik kullanmayı önerdiler.

proton, şimdiye kadarki en büyük açık deniz gemisi kadar ağırlığındadır. 3 Ana motivasyonları, kara delik motorunu bir yıldız gemisine güç sağlamak için kullanmaktı (aşağıda geri döndüğümüz bir konu), bu nedenle verimlilikten çok taşınabilirlik ile ilgileniyorlardı ve kara deliği lazer ışığıyla beslemeyi önerdiler, bu da maddeye hiçbir enerji vermiyordu. hiç dönüşüm. Onu radyasyon yerine madde ile besleyebilseniz bile, yüksek verimliliği garanti etmek zor görünüyor: protonların büyüklüklerinin binde biri kadar bir kara deliğe girmesini sağlamak için,

Büyük Hadron Çarpıştırıcısı kadar güçlü bir makine, enerjilerini artırıyor mc^2

en az bin kat daha fazla kinetik (hareket) enerjiye sahip. Kara delik buharlaştığında bu kinetik enerjinin en az% 10'u gravitonlar tarafından kaybedileceğinden, bu nedenle kara deliğe çıkarabileceğimizden ve çalıştırabileceğimizden daha fazla enerji harcıyor oluruz ve bu da negatif verimlilikle sonuçlanır. . Bir kara delik enerji santralinin umutlarını daha da karıştıran şey, hesaplamalarımızı dayandıracağımız titiz bir kuantum yerçekimi teorisinden hala yoksun olmamızdır - ancak bu belirsizlik, elbette, henüz keşfedilmemiş yeni yararlı kuantum yerçekimi etkilerinin olduğu anlamına da gelebilir.



Şekil 6.4: Dönen bir kara deliğin dönme enerjisinin bir kısmı, bir A parçacığını kara deliğin yakınına atarak ve onu yenen bir C parçasına ve A'nın başlangıçta sahip olduğundan daha fazla enerjiyle kaçan bir B parçasına bölerek elde edilebilir. .

Dönen Kara Delikler

Neyse ki, kara delikleri enerji santralleri olarak kullanmanın kuantum yerçekimi veya diğer yetersiz anlaşılmış fizik içermeyen başka yolları da var. Örneğin, birçok mevcut kara delik, olay ufkunun ışık hızına yakın bir hızla dönmesiyle çok hızlı dönüyor ve bu dönüş enerjisi elde edilebiliyor. Bir kara deliğin olay ufku, yerçekimi çok güçlü olduğu için ışığın bile kaçamayacağı bölgedir. Şekil 6.4 olay ufkunun dışında, dönen bir kara deliğin *ergosfer*

Dönen kara deliğin uzayı o kadar hızlı sürüklediği yerde, bir parçacığın hareketsiz durması ve sürüklenmemesi imkansız. Bir nesneyi ergosfere fırlatırsanız, bu nedenle deliğin etrafında dönme hızını alacaktır. Ne yazık ki, yakında kara delik tarafından yenecek, olay ufkunda sonsuza kadar kaybolacak, bu yüzden enerji çıkarmaya çalışıyorsanız bu size bir fayda sağlamaz. Ancak Roger Penrose, nesneyi akıllı bir açıyla fırlatırsanız ve onu iki parçaya ayırırsanız keşfetti. şekil 6.4 Örnekler, o zaman sadece bir parçanın yenmesini, diğerinin kara delikten başladığınızdan daha fazla enerjiyle kaçmasını sağlayabilirsiniz. Başka bir deyişle, kara deliğin dönme enerjisinin bir kısmını, işe koyabileceğiniz yararlı enerjiye başarıyla dönüştürdünüz. Bu işlemi defalarca tekrarlayarak kara deliği sağabilirsiniz. *herşey* dönme enerjisi böylece dönmeyi durdurur ve ergosferi kaybolur. Başlangıçtaki kara delik, olay ufku esasen ışık hızında hareket ederek, doğanın izin verdiği kadar hızlı dönüyorsa, bu strateji, kütlesinin% 29'unu enerjiye dönüştürmenize olanak tanır. Gece gökyüzünde kara deliklerin ne kadar hızlı döndüğü konusunda hala önemli bir belirsizlik var, ancak en iyi çalışılanların çoğu oldukça hızlı dönüyor gibi görünüyor: izin verilen maksimum değerin% 30 ila% 100'ü. Gökadamızın ortasındaki (Güneşimizin dört milyon katı ağırlığındaki) canavar kara delik dönüyor gibi görünüyor, bu nedenle kütlesinin yalnızca% 10'u faydalı enerjiye dönüştürülebilse bile, bu 400.000 güneşin aynısını verecektir. % 100 verimlilikle veya milyarlarca yıl boyunca yaklaşık 500 milyon güneş civarında Dyson kürelerinden elde ettiğimiz kadar enerjiye dönüştürüldü.

Kuasarlar

Bir başka ilginç strateji de kara deliğin kendisinden değil, içine düşen maddeden enerji elde etmektir. Doğa zaten bunu kendi başına yapmanın bir yolunu buldu: kuasar. Gaz bir kara deliğe daha da yaklaştıkça, en iç kısımları yavaş yavaş yuvarlanan pizza şeklindeki bir disk oluşturdukça, aşırı derecede ısınır ve bol miktarda radyasyon yayar. Gaz deliğe doğru aşağıya doğru düştüğünde hızlanır ve tıpkı bir paraşütçü gibi, yerçekimsel potansiyel enerjisini hareket enerjisine dönüştürür. Karmaşık türbülans, tek tek atomlar yüksek hızlarda birbirleriyle çarpışmaya başlayana kadar, gaz bloğunun koordineli hareketini giderek daha küçük ölçeklerde rastgele harekete dönüştürdüğünden hareket giderek daha karmaşık hale geliyor - bu tür rastgele harekete sahip olmak tam olarak sıcak olmak demektir. ve bu şiddetli çarpışmalar hareket enerjisini radyasyona dönüştürür. Tüm kara deliğin etrafına güvenli bir mesafede bir Dyson küresi inşa ederek, bu radyasyon enerjisi yakalanabilir ve kullanılabilir. Kara delik ne kadar hızlı dönerse, bu süreç o kadar verimli olur ve maksimum dönüşle

% 42'lik devasa bir verimlilikle enerji sağlayan kara delik. * 4 Bir yıldız kadar ağırlığa sahip kara delikler için enerjinin çoğu X-ışınları olarak ortaya çıkarken, galaksilerin merkezlerinde bulunan süper kütleli tür için çoğu kızılötesi, görünür ve morötesi ışık aralığında bir yerde ortaya çıkar.

Kara deliğinizi beslemek için yakıtınız bittikten sonra, yukarıda tartıştığımız gibi dönme enerjisini çıkarmaya geçebilirsiniz. * 5 Aslında doğa, Blandford-Znajek mekanizması olarak bilinen manyetik bir işlemle biriken gazdan gelen radyasyonu artırarak bunu kısmen yapmanın bir yolunu zaten bulmuştur. Manyetik alanların veya diğer bileşenlerin akıllıca kullanılmasıyla enerji ekstraksiyon verimliliğini% 42'nin ötesinde iyileştirmek için teknolojiyi kullanmak pekala mümkün olabilir.

Sfalerin

Maddeyi enerjiye dönüştürmenin kara delikler içermeyen bilinen başka bir yolu daha var: *Sphaleron* süreç. Kuarkları yok edebilir ve onları leptonlara dönüştürebilir: elektronlar, daha ağır kuzenleri müon ve tau parçacıkları, nötrinolar veya

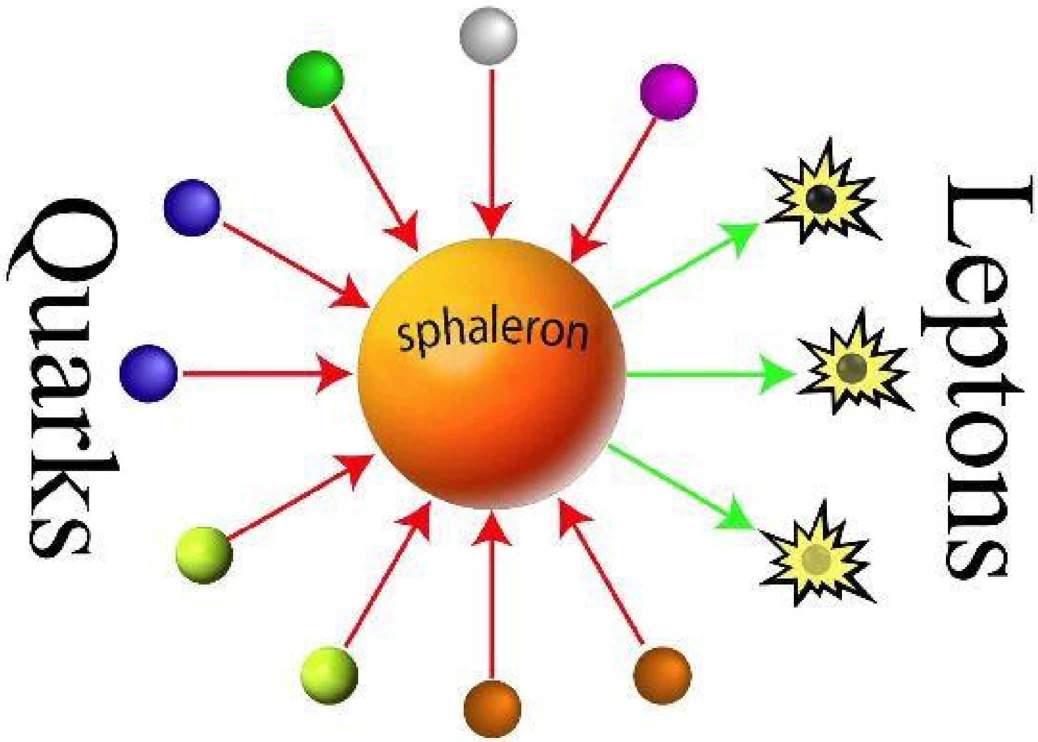
antiparçacıkları. [4] Gösterildiği gibi şekil 6.5 Parçacık fiziğinin standart modeli, uygun tada ve dönüşe sahip dokuz kuarkın bir araya gelip sphaleron adı verilen bir ara durum aracılığıyla üç leptona dönüşebileceğini öngörür. Girdi, çıktıdan daha ağır olduğu için, kütle farkı

Einstein'a göre enerjiye dönüştürülür $E = mc^2$ formül.

Bu nedenle gelecekteki akıllı yaşam, benim diyeceğim şeyi inşa edebilir. *sphalerizer:* steroidler üzerinde dizel motor gibi davranan bir enerji jeneratörü. Geleneksel bir dizel motor, hava ve dizel yağ karışımını, sıcaklık kendiliğinden tutuşması ve yanması için yeterince yüksek olana kadar sıkıştırır, ardından sıcak karışım yeniden genişler ve örneğin bir pistonu itmek gibi işlemde faydalı işler yapar. Karbondioksit ve diğer yanma gazlarının ağırlığı yaklaşık

Başlangıçta pistonda olandan% 0,00000005 daha azdır ve bu kütle farkı motoru çalıştıran ısı enerjisine dönüşür. Bir süngerleştirici, sıradan maddeyi birkaç katrilyon dereceye kadar sıkıştırır ve sonra yeniden genişlemesine izin verir.

sfaleronlar işlerini yaptıktan sonra havalıydı. * [6] Bu deneyin sonucunu zaten biliyoruz, çünkü erken Evrenimiz bunu bizim için yaklaşık 13,8 milyar yıl önce, o kadar sıcakken gerçekleştirdi: Maddenin neredeyse% 100'ü enerjiye dönüştürülür ve kalan parçacıkların milyarda birinden daha azı kalmıştır. olağan maddenin yapıldığı şeyler: kuarklar ve elektronlar. Yani tıpkı bir dizel motor gibi, bir milyardan fazla kat daha verimli! Diğer bir avantaj da, onu neyle besleyeceğiniz konusunda titiz olmanıza gerek olmamasıdır - kuarklardan yapılan herhangi bir şeyle çalışır, yani normal bir maddedir.



Şekil 6.5: Parçacık fiziğinin standart modeline göre, uygun tada ve dönüşe sahip dokuz kuark bir araya gelerek sphaleron adı verilen bir ara durum aracılığıyla üç leptona dönüşebilir. Kuarkların birleşik kütlesi (onlara eşlik eden gluon parçacıklarının enerjisi ile birlikte) leptonların kütlesinden çok daha büyüktür, bu nedenle bu işlem parlamalarla gösterilen enerjiyi serbest bırakacaktır.

Bu yüksek sıcaklık süreçleri nedeniyle, bebek Evrenimiz maddeden (daha sonra atomlara toplanan kuarklar ve elektronlar) trilyon kat daha fazla radyasyon (fotonlar ve nötrinolar) üretti. O zamandan bu yana geçen 13,8 milyar yıl boyunca, atomların galaksiler, yıldızlar ve gezegenlerde yoğunlaştığı, çoğu fotonun galaksiler arası boşlukta kaldığı ve Evrenimizin bebek resimlerini yapmak için kullanılan kozmik mikrodalga arka plan radyasyonunu oluşturduğu büyük bir ayrışma meydana geldi. . Bir galakside veya başka bir madde konsantrasyonunda yaşayan herhangi bir gelişmiş yaşam formu, mevcut maddenin çoğunu enerjiye geri döndürebilir, madde yüzdesini, erken Evrenimizden ortaya çıkan aynı küçük değere, içerideki bu sıcak yoğun koşulları kısaca yeniden yaratarak yeniden başlatabilir. bir süngerleştirici.

Gerçek bir süngerleştiricinin ne kadar verimli olacağını anlamak için, önemli pratik ayrıntılar üzerinde çalışılması gerekir: örneğin, fotonların ve nötrinoların önemli bir kısmının sıkıştırma aşamasında dışarı sızmasını önlemek için ne kadar büyük olması gerekir? Bununla birlikte, kesin olarak söyleyebileceğimiz şey, yaşamın geleceği için enerji beklentilerinin, mevcut teknolojimizin izin verdiğinden çarpıcı biçimde daha iyi olduğudur. Bir füzyon reaktörü inşa etmeyi bile başaramadık, ancak geleceğin teknolojisi on, hatta belki yüz kat daha iyisini yapabilir.

Daha İyi Bilgisayarlar Oluşturmak

Akşam yemeği yemek, enerji verimliliği üzerindeki fiziksel sınırdan 10 milyar kat daha kötü ise, bugünün bilgisayarları ne kadar verimli? Şimdi göreceğimiz gibi, o akşam yemeğinden bile daha kötü.

Sık sık arkadaşım ve meslektaşım Seth Lloyd'u MIT'de tartışmalı bir şekilde benim kadar deli olan tek kişi olarak takdim ediyorum. Kuantum bilgisayarlarda öncü çalışmalar yaptıktan sonra, tüm Evrenimizin bir kuantum bilgisayar olduğunu tartışan bir kitap yazmaya devam etti. İşten sonra sık sık bira içeriz ve henüz onun hakkında söyleyecek ilginç bir şeyi olmayan bir konu keşfetmedim. Örneğin, 2. bölümde bahsettiğim gibi, bilgi işlemin nihai sınırları hakkında söyleyecek çok şeyi var. 2000 tarihli ünlü bir makalede, hesaplama hızının enerji ile sınırlı olduğunu gösterdi: zaman içinde temel bir mantıksal işlem gerçekleştirme *T* ortalama enerji gerektirir $E = h/4\ T$, nerede *h* Planck olarak bilinen temel fizik miktarıdır

sabit. Bu, 1 kg'lık bir bilgisayarın en fazla 5 × 10 performans gösterebileceği anlamına gelir. 50

saniye başına işlem - bu, bu kelimeleri yazdığım bilgisayardan çok daha fazla 36 sıra daha fazla. Bölüm 2'de incelediğimiz gibi, hesaplama gücü her iki yılda bir ikiye katlanmaya devam ederse birkaç yüzyıl içinde oraya varacağız. Ayrıca 1 kg'lık bir bilgisayarın en fazla depolayabileceğini gösterdi.

10 31 dizüstü bilgisayarımdan yaklaşık bir milyar milyar kat daha iyi olan bit.

Seth, bu sınırlara gerçekten ulaşmanın süper zeki bir yaşam için bile zor olabileceğini, çünkü o 1 kg'lık nihai “bilgisayarın” hafızası bir termonükleer patlamaya veya Büyük Patlamamızın küçük bir parçasına benzeyeceğini kabul eden ilk kişi. Ancak, pratik sınırların nihai sınırlardan çok da uzak olmadığı konusunda iyimser. Aslında, mevcut kuantum bilgisayar prototipleri, atom başına bir bit depolayarak belleklerini çoktan küçülttüler ve bunu büyütmek,

yaklaşık 10 depolamak 25 bit / kg - dizüstü bilgisayarımdan trilyon kat daha iyi. Dahası, bu atomlar arasında iletişim kurmak için elektromanyetik radyasyon kullanmak

yaklaşık 5 × 10 izin 40 saniye başına işlem - CPU'mdan 31 büyüklük sırası daha iyi.

Özetle, gelecekteki yaşamın bir şeyleri hesaplama ve çözme potansiyeli gerçekten akıllara durgunluk vericidir: büyüklük dereceleri açısından, günümüzün en iyi süper bilgisayarları nihai 1 kg bilgisayardan, bir arabadaki yanıp sönen dönüş sinyalinden çok daha uzaktadır. , yalnızca bir parçasını depolayan bir cihaz

bilgileri saniyede bir kez açıp kapatarak.

Diğer kaynaklar

Fizik perspektifinden, gelecekteki yaşamın yaratmak isteyebileceği her şey - habitatlardan ve makinelerden yeni yaşam formlarına kadar - sadece belirli bir şekilde düzenlenmiş temel parçacıklardır. Tıpkı mavi bir balinanın krili yeniden düzenlenmesi ve krilin planktonun yeniden düzenlenmesi gibi, tüm Güneş Sistemimiz de 13.8 milyar yıllık kozmik evrim sırasında hidrojendir: yerçekimi hidrojeni, hidrojeni daha ağır atomlara yeniden düzenleyen yıldızlara yeniden düzenledi, ardından yerçekimi bu tür atomları yeniden düzenledi. kimyasal ve biyolojik süreçlerin onları hayata döndürdüğü gezegenimiz.

Teknolojik sınırına ulaşan gelecek yaşam, önce en verimli yöntemi bulmak için hesaplama gücünü kullanarak ve ardından madde yeniden düzenleme sürecini güçlendirmek için mevcut enerjisini kullanarak bu tür parçacık yeniden düzenlemelerini daha hızlı ve verimli bir şekilde gerçekleştirebilir. Maddenin hem bilgisayara hem de enerjiye nasıl dönüştürülebileceğini gördük, bu yüzden bir anlamda tek temel kaynak

gerekli. * 7 Gelecekteki yaşam, kendi maddesiyle yapabileceklerinin fiziksel sınırlarını aştığında, daha fazlasını yapmasının tek bir yolu vardır: daha fazla madde elde etmek. Ve bunu yapabilmesinin tek yolu Evrenimize genişlemektir. Uzay gemisi ho!

Kozmik Yerleşim Yoluyla Kaynak Elde Etme

Kozmik bağışımız ne kadar büyük? Spesifik olarak, fizik yasaları, yaşamın nihayetinde kullanabileceği madde miktarına hangi üst sınırlar koyar? Kozmik bağışımız elbette akıllara durgunluk verecek kadar büyük, ama tam olarak ne kadar büyük? Tablo 6.2 bazı önemli numaraları listeler. Gezegenimiz, maddesinin bu kısmının biyosferimizin bir parçası olmaması ve yerçekimi ve manyetik alan sağlamaktan başka yaşam için yararlı hiçbir şey yapmaması anlamında şu anda% 99,999999 ölüdür. Bu, yaşamı aktif olarak desteklemek için yüz milyon kat daha fazla madde kullanma potansiyelini yükseltir. Güneş Sistemimizdeki (Güneş dahil) tüm maddeyi optimum kullanıma koyabilirsek, milyonlarca kez daha iyi yapacağız. Galaksimize yerleşme kaynaklarımızı trilyonlarca kez daha büyütür.

Ne kadar uzağa gidebilirsin?

Yeterince sabırlı olursak, istediğimiz kadar çok sayıda galaksi yerleştirerek sınırsız kaynak elde edebileceğimizi düşünebilirsiniz, ancak modern kozmolojinin önerdiği bu değil! Evet, uzayın kendisi sonsuz olabilir ve sonsuz sayıda galaksi, yıldız ve gezegen içerebilir - aslında, en basit versiyonları tarafından tahmin edilen şey budur. *şişirme,* 13,8 milyar yıl önce Büyük Patlamamızı yaratan şu anda en popüler bilimsel paradigma. Bununla birlikte, sonsuz sayıda galaksi olsa bile, bunların yalnızca sınırlı bir sayısını görebiliyoruz ve onlara ulaşabiliyoruz: Yaklaşık 200 milyar galaksi görebilir ve en fazla on milyara yerleşebiliriz.

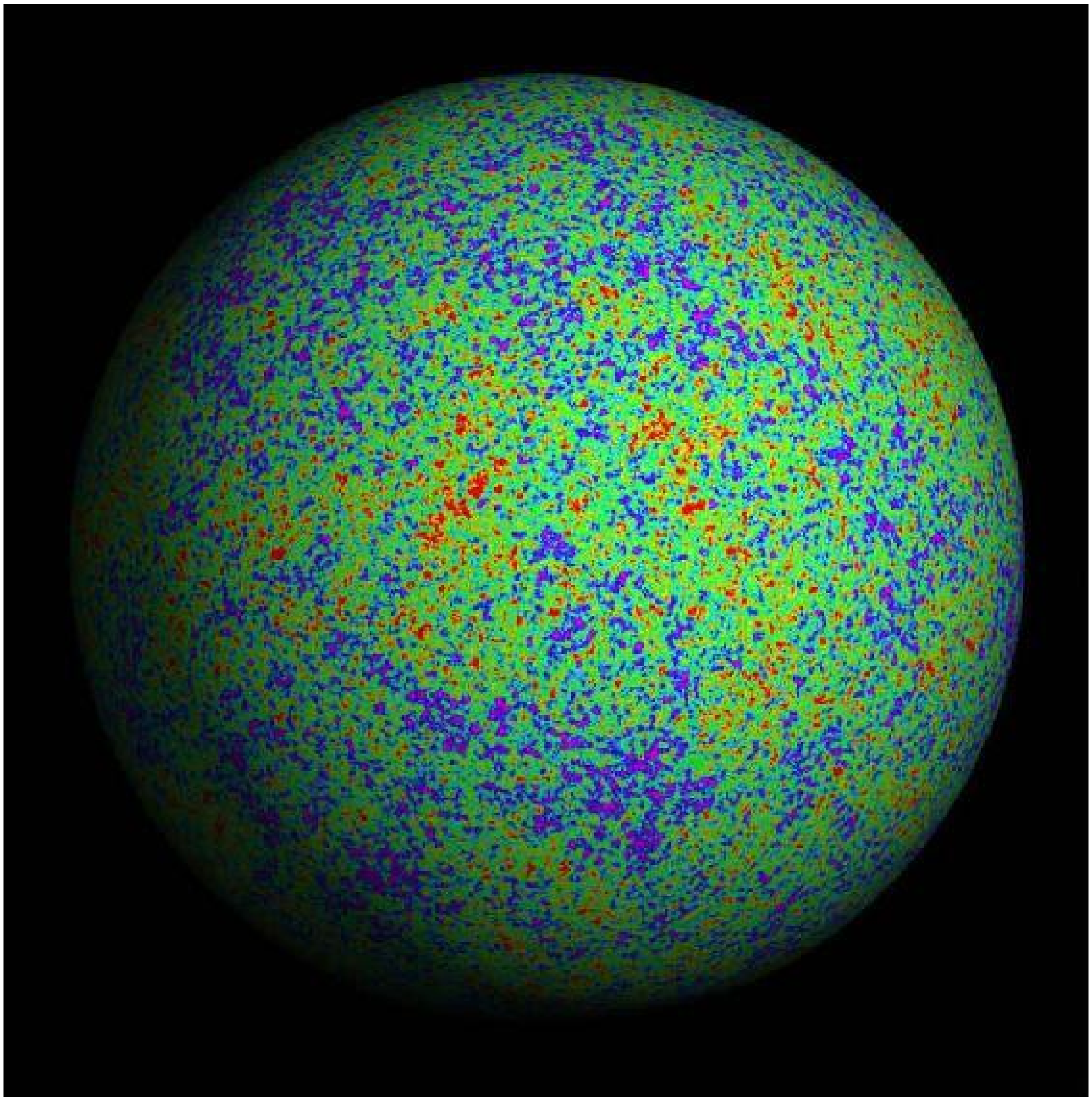
Bölge	**Parçacıklar**
Biyosferimiz	10^{43}
Bizim gezegenimiz	10^{51}
Güneş sistemimiz	10^{57}
Galaksimiz	10^{69}
Işık hızının yarısında hareket eden menzilimiz Işık	10^{75}
hızında seyahat eden menzilimiz Evrenimiz	10^{76}
	10^{78}

Tablo 6.2: Gelecekteki yaşamın kullanmak isteyebileceği yaklaşık madde parçacığı (proton ve nötron) sayısı.

Bizi sınırlandıran ışık hızıdır: yılda bir ışık yılı (yaklaşık on trilyon kilometre). Şekil 6.6 Büyük Patlamamızdan bu yana geçen 13,8 milyar yıl boyunca ışığın bize ulaştığı uzayın, "gözlemlenebilir Evrenimiz" olarak bilinen küresel bir bölge veya kısaca *"Evrenimiz."* Bile

uzay sonsuzdur, Evrenimiz sonludur, "sadece" yaklaşık 10 tane içerir ^{78} atomlar. Dahası, Evrenimizin yaklaşık% 98'i, onu görebildiğimiz ama sonsuza kadar ışık hızında seyahat etsek bile ona asla ulaşamayacağımız anlamında “gör ama dokunma” dır. Bu neden? Sonuçta, ne kadar uzağı görebileceğimizin sınırı basitçe

Evrenimiz sonsuz derecede eski değil, bu yüzden uzak ışığın bize ulaşması için henüz zamanı olmadı. Öyleyse, yolda ne kadar zaman geçirebileceğimiz konusunda bir sınırımız yoksa, gelişigüzel uzak galaksilere seyahat etmemiz gerekmez mi?



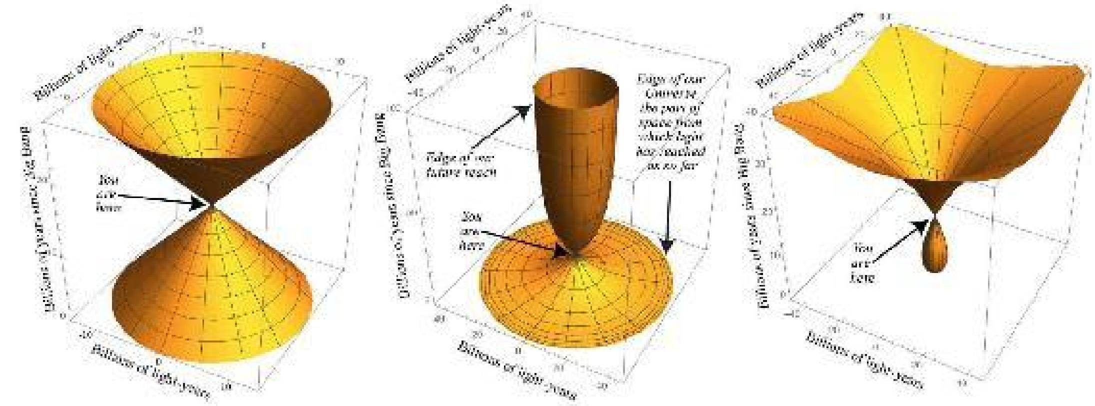
Şekil 6.6: Evrenimiz, yani Büyük Patlamamızdan bu yana 13,8 milyar yıl boyunca ışığın bize (merkezde) ulaşmak için zamanının olduğu uzayın küresel bölgesi. Desenler, Evrenimizin Planck uydusu tarafından çekilmiş bebek resimlerini gösteriyor ve sadece
400.000 yaşında, neredeyse Güneş'in yüzeyi kadar sıcak olan sıcak plazmadan oluşuyordu. Uzay muhtemelen bu bölgenin ötesinde devam ediyor ve her yıl yeni maddeler ortaya çıkıyor.

İlk zorluk, Evrenimizin genişlemesi, bu da neredeyse tüm galaksilerin bizden uzaklaştığı anlamına geliyor, bu nedenle uzak galaksilere yerleşmek bir yakalama oyunu anlamına geliyor. İkinci zorluk ise, bu kozmik genişlemenin hızlanıyor olmasıdır, çünkü bizim gücümüzün yaklaşık% 70'ini oluşturan gizemli karanlık enerji

Evren. Bunun nasıl sorun yarattığını anlamak için, bir tren platformuna girdiğinizi ve treninizin sizden uzaklaşarak yavaşça hızlandığını, ancak bir kapının davetkar bir şekilde açık bırakıldığını hayal edin. Hızlı ve aptalsan, treni yakalayabilir misin? Sonunda koşabileceğinizden daha hızlı gideceğinden, cevap açıkça trenin başlangıçta sizden ne kadar uzakta olduğuna bağlıdır: belirli bir kritik mesafenin ötesindeyse, asla yetişemezsiniz. Bizden uzaklaşmakta olan uzak galaksileri yakalamaya çalışırken aynı durumla karşı karşıyayız: Işık hızında seyahat edebilseydik bile, yaklaşık 17 milyar ışıkyılının ötesindeki tüm galaksiler sonsuza kadar ulaşılamayacak durumda kalıyor - ve bu, Evrenimizdeki galaksiler.

Ama durun: Einstein'ın özel görelilik teorisi hiçbir şeyin ışıktan hızlı gidemeyeceğini söylememiş miydi? Öyleyse galaksiler ışık hızında hareket eden bir şeyi nasıl geride bırakabilir? Cevap, özel göreliliğin yerini, hız sınırının daha liberal olduğu Einstein'ın genel görelilik teorisinin almasıdır: hiçbir şey ışık hızından daha hızlı gidemez *uzayda,* ancak alan istediği kadar hızlı genişlemekte özgürdür. Einstein ayrıca, zamanı dördüncü boyut olarak görerek bu hız sınırlarını görselleştirmenin güzel bir yolunu da verdi. *boş zaman (* görmek şekil 6.7 , üç uzay boyutundan birini çıkararak şeyleri üç boyutlu tuttuğum yer). Uzay genişlemeseydi, ışık ışınları uzay-zaman boyunca 45 derecelik eğimli çizgiler oluşturacaktı, böylece buradan ve şimdi görebildiğimiz ve ulaşabildiğimiz bölgeler koniler. Geçmişteki ışık konimiz 13.8 milyar yıl önce Büyük Patlamamızla kesilirken, gelecekteki ışık konimiz sonsuza kadar genişleyerek bize sınırsız bir kozmik donanıma erişim sağlayacaktı. Buna karşılık, şeklin orta paneli, karanlık enerjiye sahip genişleyen bir evrenin (içinde yaşadığımız Evren gibi görünüyor) ışık konilerimizi şampanya bardağı şekline deforme ettiğini ve yerleşebileceğimiz galaksi sayısını sonsuza kadar sınırladığını gösteriyor. milyar.

Bu sınır size kozmik klostrofobi hissettiriyorsa, olası bir boşlukla sizi neşelendirmeme izin verin: Hesaplamalarım, karanlık enerjinin zaman içinde sabit kaldığını ve en son ölçümlerin önerdikleriyle tutarlı olduğunu varsayıyor. Bununla birlikte, karanlık enerjinin gerçekte ne olduğu konusunda hala hiçbir fikrimiz yok, bu da karanlık enerjinin sonunda çürüyeceğine dair bir umut ışığı bırakıyor (kozmik enflasyonu açıklamak için varsayılan benzer karanlık enerji benzeri maddeye çok benzer) ve bu olursa, hızlanma yol verecek *yavaşlama,* potansiyel olarak gelecekteki yaşam formlarının, sürdükleri süre boyunca yeni galaksileri yerleştirmeye devam etmelerini sağladı.



Şekil 6.7: Bir uzay-zaman diyagramında bir olay, sırasıyla yatay ve dikey konumlarının nerede ve ne zaman meydana geldiğini kodlayan bir noktadır. Uzay genişlemiyorsa (sol panel), o zaman iki koni, Dünya'da (tepede) etkilenebileceğimiz (alt koni) uzay-zaman bölümlerini sınırlar ve (üst koni) üzerinde bir etkiye sahip olabilir, çünkü nedensel etkiler Yılda bir ışık yılı uzaklıkta olan ışıktan daha hızlı seyahat eder. Alan genişlediğinde işler daha ilginç hale gelir (sağ paneller). Standart kozmoloji modeline göre, uzay sonsuz olsa bile yalnızca uzay-zamanın sınırlı bir bölümünü görebilir ve ulaşabiliriz. Ortadaki resimde, bir şampanya kadehini andıran, uzayın genişlemesini gizleyen koordinatlar kullanıyoruz, böylece uzak galaksilerin zaman içindeki hareketleri dikey çizgilere karşılık geliyor. Şu anki bakış açımızdan, 13. Büyük Patlamamızdan 8 milyar yıl sonra, ışık ışınlarının bize sadece şampanya kadehinin tabanından ulaşma zamanı oldu ve ışık hızında seyahat etsek bile, bardağın üst kısmının dışındaki bölgelere asla ulaşamıyoruz. yaklaşık 10 milyar galaksi içerir. Sağdaki resimde, bir çiçeğin altındaki bir su damlasını anımsatan, uzayın genişlediğinin görüldüğü tanıdık koordinatları kullanıyoruz. Bu, cam tabanı bir damlacık şekline deforme eder, çünkü görebildiğimiz şeyin kenarlarındaki bölgelerin hepsi erken bir zamanda birbirine çok yakındı. uzayın genişlemesinin görüldüğü tanıdık koordinatları kullanırız. Bu, cam tabanı bir damlacık şekline deforme eder, çünkü görebildiğimiz şeyin kenarlarındaki bölgelerin hepsi erken bir zamanda birbirine çok yakındı. uzayın genişlemesinin görüldüğü tanıdık koordinatları kullanırız. Bu, cam tabanı bir damlacık şekline deforme eder, çünkü görebildiğimiz şeyin kenarlarındaki bölgelerin hepsi erken bir zamanda birbirine çok yakındı.

Ne Kadar Hızlı Gidebilirsin?

Yukarıda, bir medeniyetin her yöne ışık hızında genişlemesi durumunda kaç galaksi yerleşebileceğini keşfettik. Genel görelilik, uzayda ışık hızında roket göndermenin imkansız olduğunu söylüyor, çünkü bu,

sonsuz enerji, peki roketler pratikte ne kadar hızlı gidebilir? * [8]

NASA'nın Yeni Ufuklar roketi, 2006 yılında Plüton'a saatte yaklaşık 100.000 mil (saniyede 45 kilometre) hızla fırladığında hız rekorunu kırdı ve NASA'nın 2018 Solar Probe Plus, çok yakın düşerek dört kat daha hızlı gitmeyi hedefliyor. Güneş, ama bu bile ışık hızının% 0,1'inden daha az. Daha hızlı ve daha iyi roket arayışı, geçen yüzyılın en parlak beyinlerinden bazılarını büyüledi ve bu konuda zengin ve büyüleyici bir literatür var. Daha hızlı gitmek neden bu kadar zor? İki temel sorun, geleneksel roketlerin yanlarında taşıdıkları yakıtı hızlandırmak için yakıtlarının çoğunu harcaması ve bugünün roket yakıtının umutsuzca verimsiz olmasıdır - enerjiye dönüşen kütlesinin oranı% 0.00000005'ten çok daha iyi değildir. içinde gördüğümüz benzin tablo 6.1 . Bariz bir gelişme, daha verimli yakıta geçmektir. Örneğin, Freeman Dyson ve diğerleri, NASA'nın Orion Projesi'nde çalıştı ve 10 gün boyunca yaklaşık 300.000 nükleer bomba patlatarak ışık hızının yaklaşık% 3'üne ulaşacak kadar büyük bir uzay gemisi ile ulaşmayı hedefledi.

Asırlık bir yolculukta insanları başka bir güneş sistemine taşır. [5] Diğerleri, antimaddeyi yakıt olarak kullanmayı araştırdılar, çünkü onu sıradan maddeyle birleştirmek, neredeyse% 100 verimlilikle enerji açığa çıkardı.

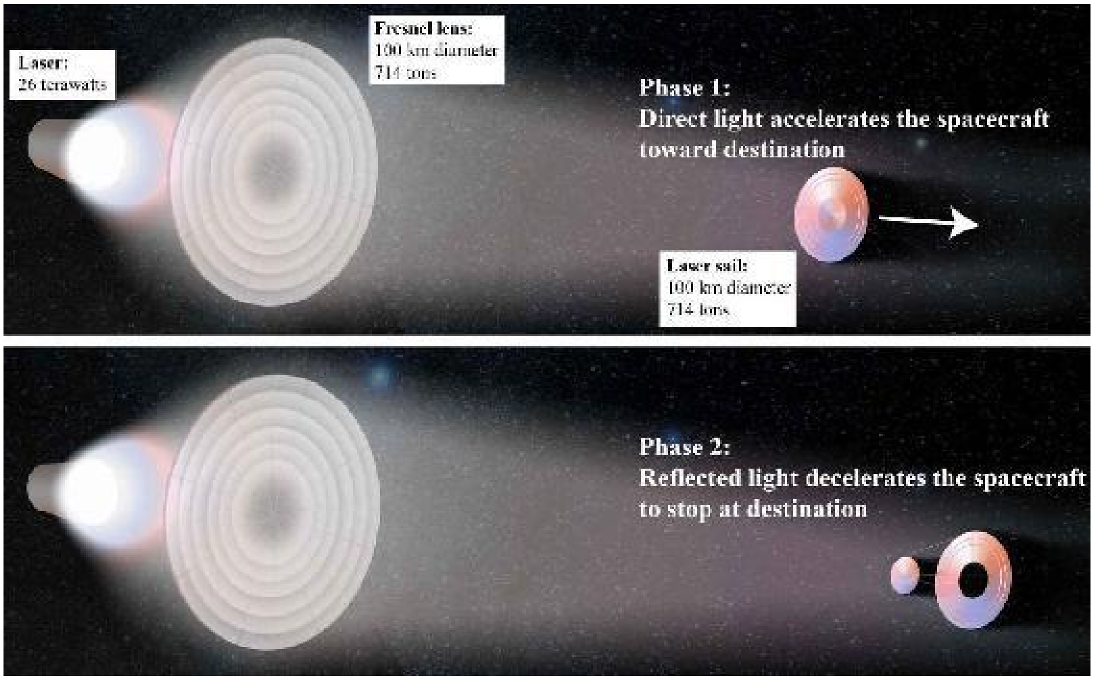
Bir diğer popüler fikir, kendi yakıtını taşıması gerekmeyen bir roket yapmaktır. Örneğin, yıldızlararası uzay mükemmel bir vakum değildir, ancak ara sıra hidrojen iyonu (yalnız bir proton: elektronunu kaybetmiş bir hidrojen atomu) içerir. 1960'da bu, fizikçi Robert Bussard'a şu anda bilinen bir şeyin arkasındaki fikri verdi.

Bussard ramjet: Bu tür iyonları yolda toplamak ve bunları yerleşik bir füzyon reaktöründe roket yakıtı olarak kullanmak. Son zamanlarda yapılan çalışmalar, bunun pratikte işe yarayıp yaramayacağına dair şüpheler uyandırsa da, yüksek teknolojili bir uzay yolculuğu uygarlığı için uygun görünen başka bir yakıtsız fikir var: lazerle yelken.

Şekil 6.8 Dyson küre yapımı için keşfettiğimiz statitleri icat eden aynı fizikçi olan Robert Forward tarafından 1984 yılında öncülüğünü yapılan zeki bir lazer yelkenli roket tasarımını göstermektedir. Bir yelkenli tekneden zıplayan hava moleküllerinin

ileri doğru itin, aynadan sıçrayan hafif parçacıklar (fotonlar) onu ileri doğru iter. Bir uzay aracına bağlı geniş bir ultra hafif yelkende güneş enerjisiyle çalışan devasa bir lazeri ışınlayarak, roketi büyük hızlara çıkarmak için kendi Güneşimizin enerjisini kullanabiliriz. Ama nasıl durursun? Forward'ın parlak makalesini okuyana kadar benden kaçan soru bu: şekil 6.8 lazer yelkeninin dış halkasının ayrıldığını ve uzay aracının önünde hareket ettiğini, lazer ışınımızı geri yansıttığını gösterir.

gemiyi ve onun daha küçük yelkenini yavaşlatın.[6] Forward, bunun insanların dört ışık yıllık yolculuğa çıkmasına izin verebileceğini hesapladı. α Centauri güneş sistemi sadece kırk yılda. Oraya vardığınızda, yeni bir dev lazer sistemi kurmayı ve Samanyolu Galaksisi boyunca yıldız atlamaya devam etmeyi hayal edebilirsiniz.



Şekil 6.8: Robert Forward'ın dört ışıkyılı uzaklıktaki α Centauri yıldız sistemine bir lazer yelken görevi tasarımı. Başlangıçta, Güneş Sistemimizdeki güçlü bir lazer, lazer yelkenine radyasyon basıncı uygulayarak uzay aracını hızlandırır. Hedefe ulaşmadan önce fren yapmak için yelkenin dış kısmı ayrılır ve lazer ışığını uzay aracına geri yansıtır.

Ama neden orada duralım? 1964'te Sovyet astronomu Nikolai Kardashev, medeniyetleri kullanabilecekleri enerjiye göre derecelendirmeyi önerdi. Kardashev ölçeğinde bir gezegenin, bir yıldızın (diyelim bir Dyson küresine sahip) ve bir galaksinin enerjisinden yararlanarak sırasıyla Tip I, Tip II ve Tip III uygarlıklarına karşılık gelir. Sonraki düşünürler, Tip IV'ün tüm erişilebilir Evrenimizi kontrol altına almaya karşılık gelmesi gerektiğini öne sürdüler. O zamandan beri hırslı yaşam formları için hem iyi hem de kötü haberler var. Kötü haber şu ki, gördüğümüz gibi, erişimimizi sınırlıyor gibi görünen karanlık enerji var. İyi haber, yapay zekanın dramatik ilerlemesi. Carl Sagan gibi iyimser vizyonerler bile, insanların diğer galaksilere ulaşma olasılıklarını oldukça umutsuz görüyorlardı. Işık hızına yakın bir hızla seyahat etsek bile milyonlarca yıl sürecek bir yolculuğun ilk yüzyılı içinde ölme eğilimimiz göz önüne alındığında. Vazgeçmeyi reddeden, donmuş astronotların yaşamlarını uzatmayı, yaşlanmalarını ışık hızına çok yakın seyahat ederek yavaşlatmayı veya bunu yapacak bir topluluk göndermeyi düşündüler.

şimdiye kadar var olan insan ırkından daha uzun olan on binlerce nesil için seyahat.

Süper zeka olasılığı bu tabloyu tamamen dönüştürerek galaksiler arası yolculuk tutkunları için çok daha umut verici hale getiriyor. Büyük insan yaşam destek sistemlerini taşıma ihtiyacını ortadan kaldıran ve yapay zeka tarafından icat edilen teknolojiyi ekleyen galaksiler arası yerleşim birdenbire oldukça basit görünüyor. Forward'ın lazer yelkeni, uzay aracının yalnızca bir "tohum sondası" içerecek kadar büyük olması gerektiğinde çok daha ucuz hale gelir: hedef güneş sisteminde bir asteroide veya gezegene inip sıfırdan yeni bir uygarlık inşa edebilen bir robot. Talimatları yanında taşıması bile gerekmez: Tek yapması gereken, ana uygarlığından ışık hızında iletilen daha ayrıntılı planları ve talimatları alacak kadar büyük bir alıcı anten inşa etmektir. Bir kez yapıldığında, galaksiye bir seferde bir güneş sistemi yerleştirmeye devam etmek için yeni tohum sondaları göndermek için yeni inşa edilmiş lazerlerini kullanıyor. Galaksiler arasındaki uçsuz bucaksız karanlık uzaylar bile, yol istasyonları olarak kullanılabilen önemli sayıda galaksiler arası yıldız (kendi galaksilerinden fırlatıldıktan sonra reddedilir) içerme eğilimindedir, böylece galaksiler arası lazer yelkenciliği için bir adadan atlama stratejisi sağlar.

Süper zeki AI tarafından başka bir güneş sistemi veya galaksi yerleştiğinde, insanları oraya getirmek kolaydır - eğer insanlar AI'nın bu amaca sahip olmasını sağlamayı başardıysa. İnsanlarla ilgili gerekli tüm bilgiler ışık hızında iletilebilir, ardından yapay zeka kuarkları ve elektronları istenen insanlara birleştirebilir. Bu, ya bir kişinin DNA'sını belirlemek için gereken iki gigabaytlık bilgiyi basitçe ileterek ve ardından AI tarafından yetiştirilecek bir bebeği kuluçkaya bırakarak oldukça düşük teknolojili yapılabilir ya da AI, kuarkları ve elektronları nano bir araya getirerek yetişkin insanlara verebilir. tüm anıların orijinallerinden Dünya'ya geri taranmasını sağlayın.

Bu, bir istihbarat patlaması varsa, asıl soru galaksiler arası yerleşimin mümkün olup olmadığı değil, ne kadar hızlı ilerleyebileceği anlamına gelir. Yukarıda araştırdığımız tüm fikirler insanlardan geldiğinden, hayatın ne kadar hızlı genişleyebileceğine dair sadece daha düşük sınırlar olarak görülmelidir; Hırslı süper zeki yaşam muhtemelen çok daha iyisini yapabilir ve zamana ve karanlık enerjiye karşı yarışta ortalama yerleşme hızındaki her% 1'lik artış kolonileşmiş galaksilerin% 3'üne dönüştüğü için sınırları zorlamak için güçlü bir teşvike sahip olacaktır.

Örneğin, bir lazer yelken sistemiyle bir sonraki yıldız sistemine 10 ışıkyılı gitmek 20 yıl sürerse ve sonra onu yerleştirmek ve yenisini inşa etmek için bir 10 yıl daha

lazerler ve tohum sondaları orada, uzayın yerleşik bölgesi, ortalama olarak ışık hızının üçte biri hızla her yöne büyüyen bir küre olacak. Amerikalı fizikçi Jay Olson, 2014'te kozmik olarak genişleyen uygarlıkların güzel ve kapsamlı bir analizinde, adadan atlamaya yüksek teknolojili bir alternatif olarak değerlendirdi.
yaklaşım, iki ayrı araştırma türü içerir: *tohum araştırmaları* ve *genişleticiler.* [7]
Tohum sondaları yavaşlayacak, yere inecek ve hedeflerine hayat verecek. Öte yandan, genişleticiler asla durmazlar: uçuş sırasında maddeyi, belki de ramjet teknolojisinin bazı gelişmiş varyantlarını kullanarak toplarlar ve bu maddeyi hem yakıt olarak hem de genişleticiler oluşturacakları hammadde olarak kullanırlar. ve kendilerinin kopyaları. Bu kendi kendini yeniden üreten genişleme filosu, yakındaki galaksilere göre her zaman sabit bir hızı (diyelim ki ışık hızının yarısı) korumak için yavaşça hızlanmaya devam edecek ve filo, kabuk başına sabit sayıda genişletici ile genişleyen küresel bir kabuk oluşturacak kadar sıklıkla çoğalacaktır. alan.

Son fakat bir o kadar da önemli olarak, yukarıdaki yöntemlerin izin verdiğinden daha hızlı genişlemeye yönelik sinsi Hail Mary yaklaşımı var: Hans Moravec'in 4. bölümdeki "kozmik spam" dolandırıcılığını kullanarak, saf, yeni evrimleşmiş medeniyetleri bir süper zeka oluşturmaya yönlendiren bir mesaj yayınlayarak Onları kaçıran makine, bir medeniyet esasen ışık hızında, baştan çıkarıcı siren şarkılarının kozmosa yayılma hızında genişleyebilir. Bu olabileceğinden

sadece ileri uygarlıkların galaksilerin çoğuna gelecekteki ışık konileriyle ulaşmaları için bir yol ve denememek için çok az teşvikleri var, dünya dışı varlıklardan gelen herhangi bir aktarımdan çok şüphelenmeliyiz! Carl Sagan'ın kitabında
İletişim, biz Dünyalılar, anlamadığımız bir makine yapmak için uzaylıların planlarını kullandık - bunu yapmayı önermiyorum ...

Özetle, kozmik yerleşimi düşünen çoğu bilim insanı ve bilim kurgu yazarı, bence süper zeka olasılığını görmezden gelme konusunda aşırı karamsar davrandılar: dikkati insan gezginlerle sınırlayarak, galaksiler arası seyahatin zorluğunu abarttılar ve dikkati teknolojiye sınırladılar. İnsanlar tarafından icat edildi, mümkün olanın fiziksel sınırlarına yaklaşmak için gereken zamanı fazla tahmin ettiler.

Kozmik Mühendislik ile Bağlı Kalmak

En son deneysel verilerin önerdiği gibi, karanlık enerji uzak galaksileri birbirinden uzaklaştırmaya devam ederse, bu, yaşamın geleceği için büyük bir sıkıntı oluşturacaktır. Bu, gelecekteki bir medeniyet bir milyon galaksiyi yerleştirmeyi başarsa bile, karanlık enerjinin on milyarlarca yıl boyunca bu kozmik imparatorluğu birbiriyle iletişim kuramayan binlerce farklı bölgeye böleceği anlamına gelir. Gelecekteki yaşam bu parçalanmayı önlemek için hiçbir şey yapmazsa, o zaman yaşamın kalan en büyük burcu, birleşik yerçekimi onları ayırmaya çalışan karanlık enerjiyi alt edecek kadar güçlü olan yaklaşık bin galaksiyi içeren kümeler olacaktır.

Süper zeki bir medeniyet bağlı kalmak istiyorsa, bu ona büyük ölçekli kozmik mühendislik yapmak için güçlü bir teşvik verecektir. Karanlık enerji onu sonsuza dek ulaşamayacağı bir yere koymadan önce, maddenin en büyük üstkümesine taşınması için ne kadar zamanı olacak? Bir yıldızı büyük mesafelere taşımak için bir yöntem, üçüncü bir yıldızı, iki yıldızın sabit bir şekilde birbirinin yörüngesinde döndüğü bir ikili sisteme dürtmektir. Romantik ilişkilerde olduğu gibi, üçüncü bir eşin tanıtılması olayları istikrarsızlaştırabilir ve üçünden birinin şiddetli bir şekilde dışarı atılmasına yol açabilir - yıldız vakasında, büyük bir hızla. Üç ortağın bazıları kara deliklerse, böylesine uçucu bir üçlü, ev sahibi galaksinin çok dışına uçacak kadar hızlı bir şekilde kütleyi fırlatmak için kullanılabilir. Ne yazık ki yıldızlara, kara deliklere veya galaksilere uygulanan bu üç cisim tekniği,

Ancak bu, açık bir şekilde süper zeki yaşamın daha iyi yöntemlerle ortaya çıkamayacağı anlamına gelmez, mesela uzaktaki galaksilerdeki kütlenin çoğunu ev kümesine gidebilecek bir uzay aracına dönüştürmek gibi. Bir süngerleştirici inşa edilebilirse, belki de maddeyi, ana kümeye ışık olarak ışınlanabilen, maddeye yeniden yapılandırılabileceği veya bir güç kaynağı olarak kullanılabileceği enerjiye dönüştürmek için bile kullanılabilir.

Nihai şans, birbirlerinden ne kadar uzakta olurlarsa olsunlar, solucan deliğinin iki ucu arasında neredeyse anlık iletişimi ve seyahat etmeyi mümkün kılan, sabit, geçilebilir solucan delikleri inşa etmenin mümkün olduğu ortaya çıkarsa olacaktır. Bir solucan deliği, araya giren uzaydan geçmeden A'dan B'ye seyahat etmenizi sağlayan uzay-zaman boyunca bir kısayoldur. Sabit solucan deliklerine Einstein'ın genel görelilik teorisi tarafından izin verilmesine ve filmlerde yer almasına rağmen

gibi *İletişim* ve *Yıldızlararası,* negatif yoğunluğa sahip tuhaf bir varsayımsal madde türünün varlığını gerektiriyorlar ve bu maddenin varlığı tam olarak anlaşılmamış kuantum yerçekimi etkilerine bağlı olabilir. Başka bir deyişle, yararlı solucan delikleri pekala imkansız hale gelebilir, ancak değilse, süper zeki yaşamın onları inşa etmek için büyük teşvikleri vardır. Solucan delikleri yalnızca tek tek galaksiler içindeki hızlı iletişimi kökten değiştirmekle kalmaz, aynı zamanda uzaktaki galaksileri merkez kümeye erken bağlayarak, gelecekteki yaşamın tüm hakimiyetinin uzun vadede bağlı kalmasına izin verir ve karanlık enerjinin iletişimi sansürleme girişimlerini tamamen engeller. İki galaksi sabit bir solucan deliği ile birbirine bağlandığında, ne kadar uzağa sürüklenirlerse düşsünler birbirlerine bağlı kalacaktır.

Eğer, kozmik mühendislikteki en iyi girişimlerine rağmen, gelecekteki bir medeniyet, bazı kısımlarının sonsuza dek temastan kopmaya mahkum olduğu sonucuna varırsa, onları bırakabilir ve iyilik dileyebilir. Bununla birlikte, bazı çok zor soruların yanıtlarını aramayı içeren iddialı hesaplama hedefleri varsa, bunun yerine bir kesme ve yakma stratejisine başvurabilir: uzaktaki galaksileri, maddelerini ve enerjilerini hesaplamaya dönüştüren devasa bilgisayarlara dönüştürebilir. karanlık enerji yanmış kalıntılarını gözden çıkarmadan önce, uzun süredir aranan cevapları anne kümesine geri iletebilecekleri umuduyla çılgın bir hız. Bu kesme ve yakma stratejisi, özellikle önceden var olan sakinleri üzecek şekilde, yalnızca “kozmik spam” yöntemiyle ulaşılabilecek kadar uzak bölgeler için uygun olacaktır.

En son kaça olur?

Uzun ömür, çoğu hırslı insan, kuruluş ve ulusun arzuladığı bir şeydir. Öyleyse, hırslı bir gelecek uygarlığı süper zeka geliştirirse ve uzun ömür istiyorsa, ne kadar sürebilir?

Uzak geleceğimizin ilk kapsamlı bilimsel analizi, Freeman Dyson tarafından gerçekleştirildi ve tablo 6.3 bazı temel bulgularını özetliyor. Sonuç şu ki, zeka müdahale etmedikçe, güneş sistemleri ve galaksiler yavaş yavaş yok edilir, sonunda her şey gelir ve sonsuza kadar solan radyasyon parıltısıyla soğuk, ölü, boş bir alan bırakmaz. Ancak Freeman, analizini iyimser bir notla bitiriyor: "Yaşam ve zekanın başarılı olma olasılığını ciddiye almak için iyi bilimsel nedenler var.

bu evrenimizi kendi amaçlarına göre şekillendirmek için. " [8]

Süper zekanın, listelenen birçok sorunu kolayca çözebileceğini düşünüyorum. tablo 6.3 , çünkü maddeyi güneş sistemleri ve galaksilerden daha iyi bir şeye yeniden düzenleyebiliyor. Nispeten düşük teknolojili bir uygarlık bile 200 milyar yıldan fazla süren düşük kütleli yıldızlara kolayca geçebileceğinden, Güneşimizin birkaç milyar yıl içinde ölmesi gibi sıkça tartışılan zorluklar çarpıcı olmayacak. Süper zeki uygarlıkların yıldızlardan daha verimli kendi enerji santrallerini inşa ettiklerini varsayarsak, aslında bunu yapmak isteyebilirler. *önlemek* Enerjiyi korumak için yıldız oluşumu: Bir yıldızın ana ömrü boyunca tüm enerji çıkışını toplamak için bir Dyson küresi kullansalar bile (toplam enerjinin yaklaşık% 0,1'ini telafi ederek), enerjinin kalan% 99,9'unun çoğunu tutamayabilirler. çok ağır yıldızlar öldüğünde boşa gitmekten. Ağır bir yıldız, enerjinin çoğunun yakalanması zor nötrinolar olarak kaçtığı bir süpernova patlamasında ölür ve çok ağır yıldızlar için, büyük miktarda kütle, bir kara delik oluşturarak boşa harcanır.

enerji 10^{67} alır dışarı sızmak için yıllar.

Ne	Ne zaman
Evrenimizin şu anki yaşı	10^{10} yıl
Karanlık enerji, galaksilerin çoğunu ulaşılamaz hale getirir Son	10^{11} yıl
yıldızlar yanar	10^{14} yıl
Yıldızlardan ayrılmış gezegenler	10^{15} yıl

Galaksilerden ayrılmış yıldızlar	10^{19} yıl
Yerçekimsel radyasyonla yörüngelerin bozulması Protonların	10^{20} yıl
bozunması (en erken)	$> 10^{34}$ yıl
Yıldız kütleli kara delikler buharlaşır Süper kütleli kara	10^{67} yıl
delikler buharlaşır Tüm maddeler bozunarak demire	10^{91} yıl
dönüşür	10^{1500} yıl
Tüm maddeler daha sonra buharlaşan kara delikler oluşturur. 1026 yıl	

Tablo 6.3: Uzak gelecek için tahminler, 2'nci ve 7'nci hariç tümü Freeman Dyson tarafından yapılmıştır. Bu hesaplamaları, karanlık enerjinin keşfedilmesinden önce yaptı, bu da birkaç tür "kozmokalipi" mümkün kılabilir.

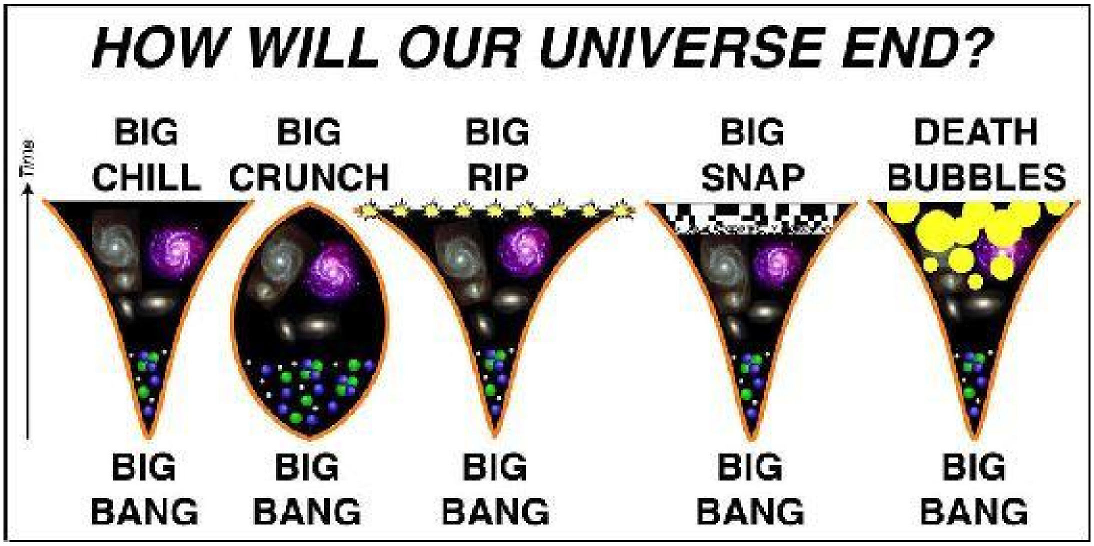
10^{10} -10^{11} yıl. Protonlar tamamen kararlı olabilir; değilse, deneyler 10'dan fazla süreceğini gösteriyor 34 yarısının çürümesi için yıllar.

Süper zeki yaşamda madde / enerji tükenmediği sürece, yaşam alanını istediği durumda tutmaya devam edebilir. Belki de sözde kullanarak protonların bozulmasını önlemenin bir yolunu bile keşfedebilir. *izlenen pot etkisi* kuantum mekaniğinin bir parçası olarak, bozunma süreci düzenli gözlemler yapılarak yavaşlatılır. Bununla birlikte, potansiyel bir gösterici vardır: *kozmokalips*

Belki bundan 10–100 milyar yıl sonra, tüm Evrenimizi yok ediyor. Karanlık enerjinin ve sicim teorisindeki ilerlemenin keşfi, Freeman Dyson'ın ufuk açıcı makalesini yazarken farkında olmadığı yeni kozmokalip senaryoları ortaya çıkardı.

Peki bundan milyarlarca yıl sonra Evrenimiz nasıl sona erecek? Yaklaşan kozmik kıyametimiz veya kozmokalipimiz için beş ana şüphelim var. şekil 6.9 : *Big Chill, Big Crunch, Big Rip, Big Snap* ve *Ölüm Balonları.* Evrenimiz şu anda yaklaşık 14 milyar yıldır genişliyor. Büyük Üşüme, Evrenimizin sonsuza kadar genişlemeye devam ettiği, kozmosumuzu soğuk, karanlık ve nihayetinde ölü bir yere sulandırdığı zamandır; bu, Freeman'ın o makaleyi yazdığı zamanki en olası sonuç olarak görüldü. Bunu TS Eliot seçeneği olarak düşünüyorum: "Dünya bu şekilde bitiyor / Bir patlama ile değil, bir sızlanma ile." Robert Frost gibi siz de dünyanın buzdan ziyade ateşle bitmesini tercih ediyorsanız, kozmik genişlemenin eninde sonunda tersine döndüğü ve her şeyin geriye doğru bir Büyük Patlama gibi felaket bir çöküşle bir araya geldiği Big Crunch için parmaklarınızı çaprazlayın. . Son olarak, Büyük Yırtılma, galaksilerimizin, gezegenlerimizin ve hatta atomlarımızın büyük bir hızla parçalandığı sabırsızlar için Büyük Üşüme gibidir.

şu andan itibaren sonlu bir zaman. Bu üçünden hangisine bahis oynamalısınız? Bu, Evrenimizin kütlesinin yaklaşık% 70'ini oluşturan karanlık enerjinin uzay genişlemeye devam ederken ne yapacağına bağlıdır. Karanlık enerjinin değişmeden yapışmasına, negatif yoğunluğa seyrelmesine veya daha yüksek yoğunluğa anti-seyreltilmesine bağlı olarak Chill, Crunch veya Rip senaryolarından herhangi biri olabilir. Hala karanlık enerjinin ne olduğu hakkında hiçbir fikrimiz olmadığı için, size nasıl bahse gireceğimi söyleyeyim:% 40 Big Chill,% 9 Big Crunch ve% 1 Big Rip.



Şekil 6.9: Evrenimizin 14 milyar yıl önce sıcak bir Büyük Patlama ile başladığını, genişlediğini ve soğuduğunu ve parçacıklarını atomlar, yıldızlar ve galaksiler halinde birleştirdiğini biliyoruz. Ama nihai kaderini bilmiyoruz. Önerilen senaryolar arasında Big Chill (ebedi genişleme), Big Crunch (recollapse), Big Rip (her şeyi parçalayan sonsuz bir genişleme oranı), Big Snap (çok fazla gerildiğinde ölümcül bir granüler yapı ortaya çıkaran uzay kumaşı), ve Death Bubbles (ışık hızında genişleyen ölümcül baloncuklarda “donma” alanı).

Paramın diğer% 50'si ne olacak? Bunu "yukarıdakilerin hiçbiri" seçeneği için saklıyorum, çünkü biz insanların alçakgönüllü olmamız ve hala anlamadığımız temel şeyler olduğunu kabul etmemiz gerektiğini düşünüyorum. Örneğin, mekanın doğası. Chill, Crunch ve Rip sonları, alanın kendisinin kararlı ve sonsuz derecede gerilebilir olduğunu varsayar. Uzayı, sadece kozmik dramanın ortaya çıktığı sıkıcı statik aşama olarak düşünürdük. Sonra Einstein bize uzayın gerçekten kilit aktörlerden biri olduğunu öğretti: Kara deliklere dönüşebilir, yerçekimi dalgaları olarak dalgalanabilir ve genişleyen bir evren olarak uzayabilir. Belki de su tenekesi gibi farklı bir aşamaya bile donabilir, yeni aşamanın hızla genişleyen ölüm balonları başka bir joker kozmokalip adayı sunar. Ölüm balonları mümkünse,

Dahası, Einstein'ın teorisi uzay genişlemesinin her zaman devam edebileceğini ve Evrenimizin Büyük Ürpertici ve Büyük

Senaryoları kopyala. Bu kulağa gerçek olamayacak kadar iyi geliyor ve öyle olduğundan şüpheleniyorum. Bir lastik bant, tıpkı boşluk gibi hoş ve kesintisiz görünür, ancak çok fazla uzatırsanız, kopar. Neden? Atomlardan oluştuğu ve yeterince gerilmesiyle kauçuğun bu granüler atomik yapısı önemli hale gelir. Uzayda da fark edemeyeceğimiz kadar küçük bir ölçekte bir tür ayrıntı düzeyi olabilir mi? Kuantum yerçekimi araştırması, bunun mantıklı olmadığını öne sürüyor.

10'dan daha küçük ölçeklerde geleneksel üç boyutlu uzay hakkında konuşun 34
metre. Eğer gerçekten dehşet verici bir “Büyük Çarpışma” ya uğramadan uzayın sonsuza kadar uzatılamayacağı doğruysa, gelecekteki medeniyetler uzayın genişlemeyen en büyük bölgesine (büyük bir galaksi kümesi) ulaşabilecekleri yere taşınmak isteyebilirler.

Ne Kadar Hesaplayabilirsiniz?

Gelecek hayatın ne kadar uzun olduğunu keşfettikten sonra *Yapabilmek* son olarak, ne kadar süreceğini keşfedelim *istemek* sona kadar. Mümkün olduğu kadar uzun yaşamak istemenin doğal olduğunu düşünseniz de, Freeman Dyson bu arzu için daha niceliksel bir argüman da verdi: yavaş hesapladığınızda hesaplamanın maliyeti düşer, böylece işleri yavaşlatırsanız sonuçta daha çok iş yaparsınız. mümkün olduğu kadar. Freeman, Evrenimiz sonsuza kadar genişlemeye ve soğumaya devam ederse, sonsuz miktarda hesaplamanın mümkün olabileceğini bile hesapladı.

Yavaşlık ille de sıkıcı anlamına gelmez: eğer gelecekteki yaşam simüle edilmiş bir dünyada yaşıyorsa, öznel olarak deneyimlenen zaman akışının, simülasyonun dış dünyada yürütüldüğü buzul hızı ile hiçbir ilgisi yoktur, bu nedenle sonsuz olasılıkları hesaplama, simüle edilmiş yaşam formları için öznel ölümsüzlüğe dönüşebilir. Kozmolog Frank Tipler, sıcaklık ve yoğunluk artarken hesaplamaları sonsuzluğa doğru hızlandırarak, Big Crunch'tan önceki son anlarda öznel ölümsüzlük elde edebileceğinizi tahmin etmek için bu fikir üzerine inşa etti.

Karanlık enerji, hem Freeman'ın hem de Frank'in sonsuz hesaplama hayallerini bozuyor gibi göründüğünden, gelecekteki süper zeka, enerji kaynaklarını nispeten hızlı bir şekilde tüketmeyi, kozmik ufuklar ve proton bozunması gibi sorunlara girmeden önce bunları hesaplamalara dönüştürmeyi tercih edebilir. Toplam hesaplamayı en üst düzeye çıkarmak nihai hedefse, en iyi strateji çok yavaş (yukarıda belirtilen sorunlardan kaçınmak için) ve çok hızlı (hesaplama başına gerekenden daha fazla enerji harcamak) arasında bir değiş tokuş olacaktır.

Bu bölümde keşfettiğimiz her şeyi bir araya getirdiğimizde, maksimum verimli enerji santrallerinin ve bilgisayarların süper zeki yaşamın akıllara durgunluk veren miktarda hesaplama yapmasını sağlayacağını söylüyor. Yüz yıl boyunca on üç watt'lık beyninize güç vermek, yaklaşık yarım miligram maddede enerji gerektirir - tipik bir şeker tanesinden daha az. Seth Lloyd'un çalışması, beynin katrilyon kat daha fazla enerji verimli hale getirilebileceğini ve bu şeker tanesinin şimdiye kadar yaşamış tüm insan yaşamlarının simülasyonunu ve binlerce kat daha fazla insanı güçlendirmesini sağlayabileceğini öne sürüyor. Mevcut Evrenimizdeki tüm mesele

insanları simüle etmek için kullanılabilir, bu 10'dan fazla ^{69} hayatlar - veya süper zeki yapay zekanın hesaplama gücüyle yapmayı tercih ettiği başka şeyler. Hatta daha fazla

simülasyonları daha yavaş çalıştırılsaydı hayatlar mümkün olabilirdi. [9] Tersine, kitabında *Süper zeka,* Nick Bostrom, 10^{58} enerji verimliliği hakkında daha muhafazakar varsayımlarla insan yaşamı simüle edilebilir. Bununla birlikte, bu rakamları dilimleyip parçalara ayırsak, bunlar çok büyük ve yaşamın bu gelecekteki potansiyelinin boşa harcanmamasını sağlama sorumluluğumuz gibi. Bostrom'un dediği gibi: "Böyle bir yaşamın tamamında yaşanan tüm mutluluğu tek bir damla neşeyle temsil edersek, o zaman bu ruhların mutluluğu her saniye Dünya okyanuslarını doldurup yeniden doldurabilir ve yüz milyar milyar boyunca bunu yapmaya devam edebilir. bin yıl. Bunların gerçekten sevinç gözyaşları olduğundan emin olmamız gerçekten önemli. "

Kozmik Hiyerarşiler

Işık hızı sadece yaşamın yayılmasını değil, aynı zamanda yaşamın doğasını da sınırlayarak iletişim, bilinç ve kontrol üzerinde güçlü kısıtlamalar getirir. Öyleyse, kozmosumuzun çoğu sonunda canlanırsa, bu hayat nasıl olacak?

Düşünce Hiyerarşileri

Hiç elinizle sinek atmayı denediniz ve başaramadınız mı? Sizden daha hızlı tepki verebilmesinin nedeni daha küçük olmasıdır, böylece bilginin gözleri, beyni ve kasları arasında dolaşması daha az zaman alır. Bu "daha büyük = daha yavaş" ilkesi, hız sınırının elektrik sinyallerinin nöronlardan ne kadar hızlı geçebileceğiyle belirlendiği sadece biyoloji için değil, aynı zamanda hiçbir bilgi ışıktan daha hızlı hareket edemiyorsa gelecekteki kozmik yaşam için de geçerlidir. Dolayısıyla, akıllı bir bilgi işleme sistemi için büyük olmak, ilginç bir değiş tokuş içeren karma bir lütuftur. Bir yandan, daha büyük olmak, daha karmaşık düşünceleri mümkün kılan daha fazla parçacık içermesine izin verir. Öte yandan, ilgili bilginin tüm bölümlerine yayılması artık daha uzun sürdüğü için bu, gerçekten küresel düşüncelere sahip olma hızını yavaşlatıyor.

Öyleyse hayat evrenimizi yutarsa, hangi biçimi seçecek: basit ve hızlı mı yoksa karmaşık ve yavaş mı? Dünya yaşamının yaptığı gibi aynı seçimi yapacağını tahmin ediyorum: ikisi de! Dünya'nın biyosferinin sakinleri, şaşırtıcı boyutlarda,
devasa iki yüz tonluk mavi balinalardan minyon 10^{-16} kg bakteri *Pelagibacter,* Tüm dünyadaki balıkların toplamından daha fazla biyokütle oluşturduğuna inanılıyor. Dahası, büyük, karmaşık ve yavaş olan organizmalar, basit ve hızlı olan daha küçük modüller içererek genellikle yavaşlıklarını azaltırlar. Örneğin, göz kırpma refleksiniz son derece hızlıdır, çünkü beyninizin çoğunu içermeyen küçük ve basit bir devre tarafından uygulanmaktadır: Eğer bu zor uçma yanlışlıkla gözünüze doğru yönelirse, onda bir oranında göz kırparsınız. ilgili bilginin beyninize yayılması ve sizi bilinçli olarak ne olduğunun farkına varması için zaman bulmadan çok önce. Bilgi işlemeyi bir modül hiyerarşisi halinde düzenleyerek, biyosferimiz hem hızı hem de karmaşıklığı elde ederek hem pastaya sahip olmayı hem de onu yemeyi başarır. Biz insanlar paralel hesaplamayı optimize etmek için zaten bu aynı hiyerarşik stratejiyi kullanıyoruz.

İç iletişim yavaş ve maliyetli olduğu için, gelecekteki kozmik yaşamın da aynısını yapmasını bekliyorum, böylece hesaplamalar olabildiğince yerel olarak yapılacaktır. Bir hesaplama 1 kg'lık bir bilgisayarla yapılacak kadar basitse, onu galaksi büyüklüğündeki bir bilgisayara yaymak ters etki yaratır, çünkü her hesaplama adımından sonra bilgilerin ışık hızında paylaşılmasını beklemek, yaklaşık olarak saçma sapan bir gecikmeye neden olur. Adım başına 100.000 yıl.

Varsa, gelecekteki bu bilgi işlemenin ne olacağı *bilinçli* öznel bir deneyimi içerme anlamında, 8. bölümde inceleyeceğimiz tartışmalı ve büyüleyici bir konudur. Bilinç, sistemin farklı bölümlerinin birbiriyle iletişim kurmasını gerektiriyorsa, o zaman daha büyük sistemlerin düşünceleri zorunludur. Yavaş. Siz veya gelecekteki Dünya büyüklüğünde bir süper bilgisayar saniyede birçok düşünceye sahip olabilirken, galaksi büyüklüğünde bir zihin her yüz bin yılda bir yalnızca bir düşünceye sahip olabilir ve bir milyar ışık yılı büyüklüğündeki bir kozmik aklın yalnızca 10 karanlık enerji onu bağlantısız parçalara ayırmadan önce toplam düşünceler. Öte yandan, bu birkaç değerli düşünce ve beraberindeki deneyimler oldukça derin olabilir!

Kontrol Hiyerarşileri

Düşüncenin kendisi çok çeşitli ölçeklere yayılan bir hiyerarşi içinde organize edilmişse, o zaman güç ne olacak? Bölüm 4'te, akıllı varlıkların doğal olarak kendilerini nasıl güç hiyerarşileri içinde organize ettiklerini, herhangi bir varlığın stratejilerini değiştirselerdi daha kötüye gideceklerini keşfettik. İletişim ve ulaşım teknolojisi ne kadar iyi olursa, bu hiyerarşiler o kadar büyüyebilir. Süper zeka bir gün kozmik ölçeklere genişlerse, güç hiyerarşisi nasıl olacak? Serbest ve ademi merkeziyetçi mi yoksa son derece otoriter mi olacak? İşbirliği esas olarak karşılıklı yarar mı yoksa baskı ve tehditlere mi dayanacak?

Bu sorulara ışık tutmak için, hem havucu hem de çubuğu ele alalım: Kozmik ölçeklerde işbirliği için hangi teşvikler var ve bunu uygulamak için hangi tehditler kullanılabilir?

Havuç ile Kontrol Etmek

Yeryüzünde, *Ticaret* geleneksel bir işbirliğinin itici gücü olmuştur çünkü bir şeyler üretmenin göreceli zorluğu gezegende değişiklik göstermektedir. Bir kilogram gümüş çıkarmak, bir bölgede bir kilogram bakır çıkarmaktan 300 kat daha pahalıyken, başka bir bölgede yalnızca 100 kat daha pahalıysa, her ikisi de 200 kg bakır ile 1 kg gümüş ticareti yaparak öne geçecektir. Bir bölge diğerinden çok daha yüksek teknolojiye sahipse, her ikisi de benzer şekilde yüksek teknolojili malların hammaddelere karşı ticaretinden yararlanabilir.

Bununla birlikte, süper zeka, temel parçacıkları herhangi bir madde biçimine kolayca yeniden düzenleyebilen bir teknoloji geliştirirse, uzun mesafeli ticaret için teşviklerin çoğunu ortadan kaldıracaktır. Parçacıklarını yeniden düzenleyerek bakırı gümüşe dönüştürmek daha basit ve daha hızlı iken, uzaktaki güneş sistemleri arasında gümüş taşımak neden zahmete girsin? Her iki yerde de hem bilgi birikimi hem de ham maddeler (herhangi bir sorun olacak) varken galaksiler arasında yüksek teknoloji ürünü makinelerin taşınmasına neden gerek kalsın? Tahminim, süper zekayla dolu bir kozmosta, uzun mesafeler taşımaya değecek neredeyse tek emtia olacaktır. *bilgi.* Bunun tek istisnası, kozmik mühendislik projelerinde kullanılacak madde olabilir - örneğin, yukarıda bahsedilen karanlık enerjinin medeniyetleri parçalamaya yönelik yıkıcı eğilimine karşı koymak için. Geleneksel insan ticaretinin tersine, bu madde her ne olursa olsun, herhangi bir uygun toplu biçimde, hatta belki bir enerji ışını olarak gönderilebilir, çünkü alıcı süper zeka, onu istediği nesnelere hızla yeniden düzenleyebilir.

Bilgi paylaşımı veya ticareti, kozmik işbirliğinin ana itici gücü olarak ortaya çıkarsa, o zaman ne tür bilgiler söz konusu olabilir? İstenilen herhangi bir bilgi, eğer onu oluşturmak büyük ve zaman alıcı bir hesaplama çabası gerektiriyorsa değerli olacaktır. Örneğin, bir süper-zeka, fiziksel gerçekliğin doğası hakkındaki zorlu bilimsel sorulara, teoremler ve optimal algoritmalar hakkındaki zorlu matematiksel sorulara ve muhteşem teknolojinin en iyi nasıl inşa edileceğine dair zorlu mühendislik sorularına cevaplar isteyebilir. Hedonistik yaşam formları harika dijital eğlence ve simüle edilmiş deneyimler isteyebilir ve kozmik ticaret, bitcoin ruhu içinde bir tür kozmik kripto para birimi talebini artırabilir.

Bu tür paylaşım fırsatları, yalnızca kabaca eşit güce sahip varlıklar arasında bilgi akışını değil, aynı zamanda yukarı ve aşağı güçleri de teşvik edebilir.

hiyerarşiler, örneğin güneş sistemi boyutundaki düğümler ile galaktik bir merkez arasında veya galaksi boyutlu düğümler ile kozmik bir merkez arasında. Düğümler bunu daha büyük bir şeyin parçası olmanın zevki, tek başlarına geliştiremeyecekleri yanıtlar ve teknolojiler sağlaması ve dış tehditlere karşı savunma için isteyebilir. Ayrıca, yedekleme yoluyla neredeyse ölümsüzlük vaadine de değer verebilirler: tıpkı birçok insan, fiziksel bedenleri öldükten sonra zihinlerinin yaşayacağı inancıyla teselli alırken, gelişmiş bir YZ, aklının ve bilgisinin daha sonra bir merkez süper bilgisayarda yaşamasını takdir edebilir. orijinal fiziksel donanımı enerji rezervlerini tüketti.

Tersine, hub, sonuçlara acilen ihtiyaç duyulmayan devasa uzun vadeli bilgi işlem görevlerinde düğümlerinin kendisine yardımcı olmasını isteyebilir, böylece yanıtlar için binlerce veya milyonlarca yıl beklemeye değer. Yukarıda incelediğimiz gibi, merkez aynı zamanda kendi düğümlerinin galaktik kütle konsantrasyonlarını birlikte hareket ettirerek yıkıcı karanlık enerjiye karşı koymak gibi devasa kozmik mühendislik projelerini gerçekleştirmeye yardımcı olmasını isteyebilir. Çaprazlanabilir solucan deliklerinin mümkün ve inşa edilebilir olduğu ortaya çıkarsa, o zaman bir merkezin en büyük önceliği muhtemelen karanlık enerjiyi engellemek ve imparatorluğunu sonsuza kadar bağlı tutmak için bunlardan bir ağ inşa etmek olacaktır. Bir kozmik süper zekanın hangi nihai hedeflere sahip olabileceğine dair sorular, 7. bölümde daha ayrıntılı olarak inceleyeceğimiz büyüleyici ve tartışmalı bir sorudur.

Çubuk ile Kontrol Etmek

Karasal imparatorluklar genellikle astlarını hem havucu hem de sopayı kullanarak işbirliği yapmaya zorlar. Roma İmparatorluğunun tebaası, kendilerine sunulan teknoloji, altyapı ve savunmaya işbirliklerinin karşılığı olarak değer verirken, isyan ya da vergi ödememenin kaçınılmaz sonuçlarından da korkuyorlardı. Roma'dan uzaktaki eyaletlere asker göndermek için gereken uzun süre nedeniyle, gözdağı vermenin bir kısmı yerel birliklere ve neredeyse anlık cezalar verme yetkisine sahip sadık yetkililere verildi. Süper zeki bir merkez, kozmik imparatorluğu boyunca sadık muhafızlardan oluşan bir ağ kurmaya benzer bir strateji kullanabilir. Süper zeki konuları kontrol etmek zor olabileceğinden, uygulanabilir en basit strateji, nispeten aptal olmaları nedeniyle% 100 sadık olacak şekilde programlanmış AI korumaları kullanmak olabilir.

Örneğin, merkez YZ'nin kontrol etmek istediği güneş sistemi boyutundaki bir uygarlığın yakınına beyaz bir cücenin yerleştirilmesini düzenlediğini varsayalım. Beyaz cüce, mütevazı ağırlığa sahip bir yıldızın yanmış kabuğu. Büyük ölçüde karbondan oluşan, gökteki dev bir elması andırır ve o kadar kompakttır ki, Dünya'dan daha küçükken Güneş'ten daha ağır olabilir. Hintli fizikçi Subrahmanyan Chandrasekhar ünlü bir şekilde kanıtladı: *Chandrasekhar sınırı,* Güneşimizin kütlesinin yaklaşık 1,4 katı, 1A tipi süpernova olarak bilinen felaket bir termonükleer patlamaya uğrayacak. Merkez yapay zekası, bu beyaz cücenin Chandrasekhar sınırına son derece yakın olmasını acımasızca düzenlediyse, koruma YZ'si son derece aptal olsa bile etkili olabilir (aslında, büyük ölçüde aptal olduğu için): basitçe doğrulamak için programlanabilir. boyun eğdirilmiş uygarlığın aylık kozmik bitcoin kotasını, matematiksel kanıtları veya diğer vergiler öngörülen diğer vergileri teslim ettiğini ve aksi takdirde, süpernovayı ateşlemek ve tüm bölgeyi paramparça etmek için beyaz cüceye yeterince kütle attığını.

Gökada büyüklüğündeki uygarlıklar, çok sayıda kompakt nesneyi galaksi merkezindeki canavar kara deliğin etrafındaki sıkı yörüngelere yerleştirerek ve bu kütleleri örneğin çarpışarak gaza dönüştürmekle tehdit ederek benzer şekilde kontrol edilebilir. Bu gaz daha sonra kara deliği beslemeye başlayacak ve onu güçlü bir kuasara dönüştürerek potansiyel olarak galaksinin çoğunu yaşanmaz hale getirecektir.

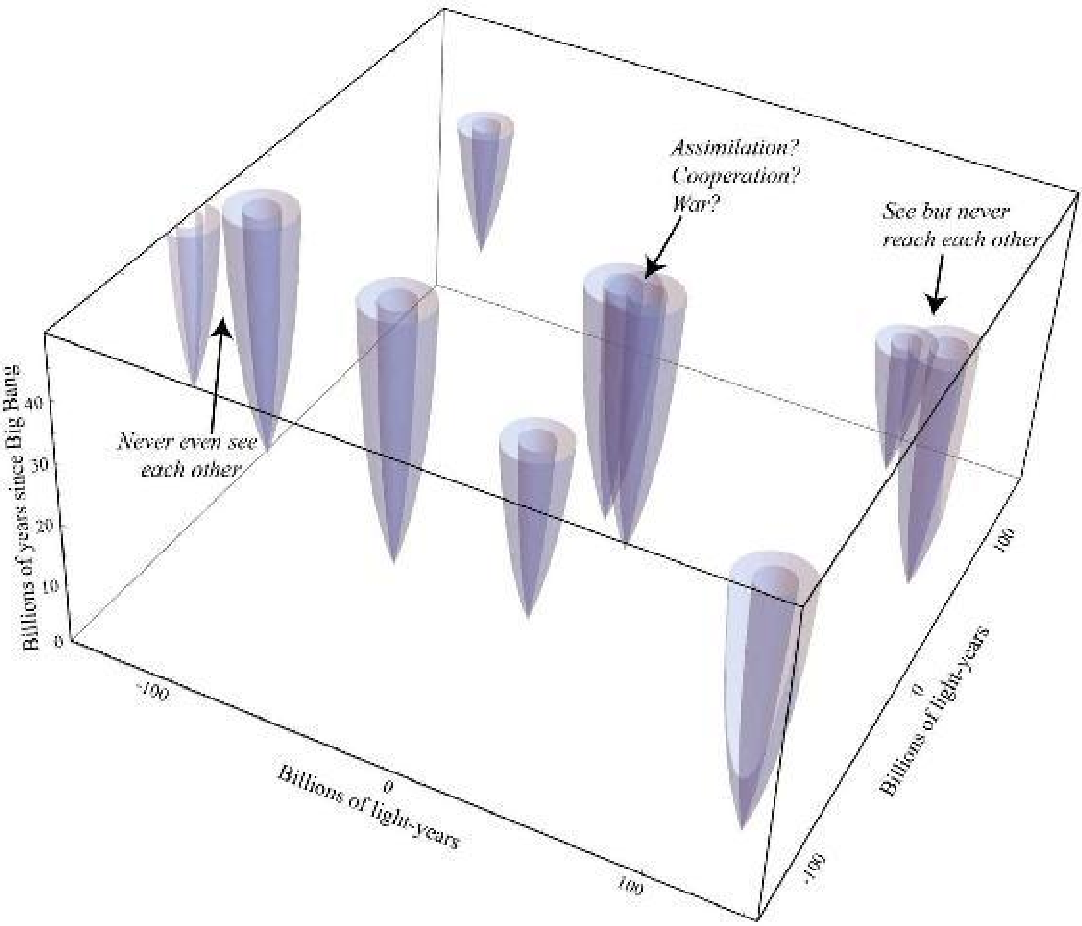
Özetle, gelecekteki yaşamın kozmik mesafelerde işbirliği yapması için güçlü teşvikler var, ancak bu tür bir işbirliğinin esas olarak karşılıklı yararlara mı yoksa acımasız tehditlere mi dayandırılacağı çok açık bir sorudur - fiziğin dayattığı sınırlar her iki senaryoya da izin veriyor gibi görünmektedir. sonuç, geçerli hedeflere ve değerlere bağlı olacaktır. Gelecekteki yaşamın bu hedeflerini ve değerlerini etkileme becerimizi 7. bölümde keşfedeceğiz.

Medeniyetler Çatıştığında

Şimdiye kadar, yaşamın tek bir zeka patlamasından kozmosumuza genişlediği senaryoları tartıştık. Peki hayat birden fazla yerde bağımsız olarak gelişirse ve genişleyen iki uygarlık bir araya gelirse ne olur?

Rastgele bir güneş sistemini düşünürseniz, yaşamın gezegenlerinden birinde gelişmesi, ileri teknoloji geliştirmesi ve uzaya genişlemesi ihtimali vardır. Teknolojik yaşam burada Güneş Sistemimizde geliştiğinden ve fizik yasaları uzayın yerleşmesine izin verdiği için bu olasılık sıfırdan büyük görünüyor. Uzay yeterince büyükse (aslında, kozmolojik enflasyon teorisi onun geniş ya da sonsuz olduğunu öne sürüyor), o zaman bu tür genişleyen pek çok uygarlık olacaktır. şekil 6.10 . Jay Olson'ın yukarıda bahsedilen makalesi, bu tür genişleyen kozmik biyosferlerin zarif bir analizini içeriyor ve Toby Ord, Future of Humanity Enstitüsü'ndeki meslektaşlarıyla benzer bir analiz gerçekleştirdi. Üç boyutta bakıldığında, bu kozmik biyosferler, medeniyetler her yöne aynı hızla genişledikleri sürece, kelimenin tam anlamıyla kürelerdir. Uzay zamanında, şampanya kadehinin üst kısmı gibi görünürler. şekil 6.7 çünkü karanlık enerji nihayetinde her uygarlığın ulaşabileceği galaksiyi sınırlar.

Yerleşen komşu medeniyetler arasındaki mesafe, karanlık enerjinin genişlemesine izin verdiğinden çok daha büyükse, o zaman birbirleriyle asla temas kurmazlar ve hatta birbirlerinin varlığını öğrenmezler, böylece yalnızmış gibi hissederler. kozmosta. Evrenimiz daha verimli ise, komşular birbirine daha yakın olsa da, bazı medeniyetler sonunda üst üste gelecektir. Bu örtüşen bölgelerde ne olur? İşbirliği mi, rekabet mi yoksa savaş mı olacak?



Şekil 6.10: Eğer yaşam uzay-zamanda (yerler ve zamanlar) birden fazla noktada bağımsız olarak gelişirse ve uzayı kolonileştirmeye başlarsa, uzay her biri şampanya kadehinin tepesine benzeyen genişleyen bir kozmik biyosferler ağı içerecektir. şekil 6.7 . Her biyosferin tabanı, kolonizasyonun başladığı yeri ve zamanı temsil eder. Opak ve yarı saydam şampanya kadehleri sırasıyla ışık hızının% 50 ve% 100'ünde kolonizasyona karşılık gelir ve örtüşmeler bağımsız medeniyetlerin buluştuğu yeri gösterir.

Avrupalılar, üstün teknolojiye sahip oldukları için Afrika ve Amerika'yı fethedebildiler. Bunun tersine, iki süper zeki medeniyetin birbiriyle karşılaşmasından çok önce, teknolojilerinin yalnızca fizik yasalarıyla sınırlı olarak aynı seviyede yaylanacağı makul. Bu, bir süper zekanın istese bile diğerini kolayca fethedebilmesini olası görünmüyor. Dahası, hedefleri görece uyumlu olacak şekilde geliştiyse, fetih veya savaş arzulamak için çok az sebepleri olabilir. Örneğin, eğer

Hem olabildiğince çok güzel teoremi kanıtlamaya hem de mümkün olduğunca akıllı algoritmalar icat etmeye çalışıyorlar, bulgularını paylaşabiliyorlar ve her ikisi de daha iyi durumda oluyor. Sonuçta, bilgi insanların genellikle uğruna mücadele ettiği kaynaklardan çok farklıdır, çünkü aynı anda hem başkasına verebilir hem de saklayabilirsiniz.

Köktendinci bir kült veya yayılan bir virüs gibi bazı genişleyen uygarlıkların esasen değişmez hedefleri olabilir. Bununla birlikte, bazı ileri uygarlıkların daha çok açık fikirli insanlara benzediği - yeterince zorlayıcı argümanlarla sunulduğunda hedeflerini ayarlamaya istekli oldukları da makul. İkisi bir araya gelirse, en ikna edici olanın galip geldiği ve hedeflerinin ışık hızıyla diğer uygarlığın kontrolündeki bölgeye yayıldığı bir silah değil fikir çatışması olacaktır. Komşularınızı asimile etmek yerleşimden daha hızlı bir genişleme stratejisidir, çünkü etki alanınız fikirlerin hareket ettiği hızda yayılabilir (telekomünikasyon kullanan ışık hızı), oysa fiziksel yerleşim kaçınılmaz olarak ışık hızından daha yavaş ilerler. *Yıldız Savaşları,* ama fikirlerin ikna edici üstünlüğüne dayanan gönüllü, asimile edilmişleri daha iyi durumda bırakıyor.

Geleceğin kozmosunun hızla genişleyen iki tür baloncuk içerebileceğini gördük: genişleyen uygarlıklar ve ışık hızında genişleyen ve tüm temel parçacıklarımızı yok ederek uzayı yaşanmaz hale getiren ölüm balonları. Hırslı bir medeniyet böylece üç tür bölgeyle karşılaşabilir: ıssız bölgeler, yaşam balonları ve ölüm balonları. İşbirliği yapmayan rakip medeniyetlerden korkuyorsa, hızlı bir "toprak gaspı" başlatmak ve ıssız bölgeleri rakiplerinden önce yerleşmek için güçlü bir teşviki var. Bununla birlikte, başka medeniyetler olmasa bile aynı yayılmacı teşvike sahiptir, sadece karanlık enerji onları ulaşılmaz hale getirmeden önce kaynakları elde etmek. Başka bir genişleyen medeniyete çarpmanın, ıssız bir alana çarpmaktan daha iyi veya daha kötü olabileceğini gördük. bu komşunun ne kadar işbirlikçi ve açık fikirli olduğuna bağlı. Bununla birlikte, herhangi bir yayılmacı uygarlığa (uygarlığınızı ataçlara dönüştürmeye çalışan biri bile) çarpmak, onunla savaşmaya çalışsanız da ya da onunla mantık yürütseniz de ışık hızında genişlemeye devam edecek bir ölüm balonundan daha iyidir. Ölüm kabarcıklarına karşı tek korumamız, uzaktaki insanların bize ulaşmasını engelleyen karanlık enerjidir. Öyleyse, eğer ölüm balonları gerçekten yaygınsa, o zaman karanlık enerji aslında bizim düşmanımız değil dostumuzdur. uzak olanların bize ulaşmasını engelleyen. Öyleyse, eğer ölüm balonları gerçekten yaygınsa, o zaman karanlık enerji aslında bizim düşmanımız değil dostumuzdur. uzak olanların bize ulaşmasını engelleyen. Öyleyse, eğer ölüm balonları gerçekten yaygınsa, o zaman karanlık enerji aslında bizim düşmanımız değil dostumuzdur.

Yalnız mıyız?

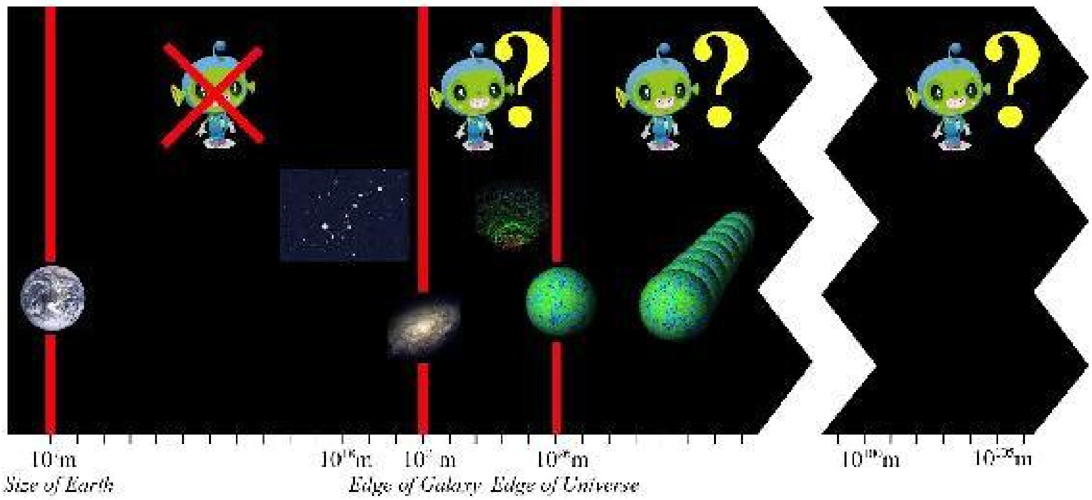
Pek çok insan Evrenimizin çoğunda ileri yaşam olduğunu varsayar, böylece insan neslinin tükenmesi kozmik bir bakış açısıyla pek önemli olmaz. Sonuçta, bazı ilham verici ise neden kendimizi yok etme konusunda endişelenelim? *Yıldız Savaşları* –Medeniyet gibi, yakında Güneş Sistemimizi yaşamla dolduracak ve yeniden tohumlayacak, hatta belki de onların ileri teknolojilerini bizi yeniden inşa etmek ve diriltmek için kullanıyor? Bunu görüyorum *Yıldız Savaşları* Tehlikeli bir varsayım, çünkü bizi sahte bir güvenlik duygusuna sürükleyebilir ve medeniyetimizi kayıtsız ve umursamaz hale getirebilir. Aslında, Evrenimizde yalnız olmadığımıza dair bu varsayımın sadece tehlikeli değil aynı zamanda muhtemelen yanlış olduğunu düşünüyorum.

Bu bir azınlık görüşü, * 9 ve ben yanılıyor olabilirim, ama en azından şu anda göz ardı edemeyeceğimiz bir olasılık, bu da bize onu güvenli bir şekilde oynamamız ve uygarlığımızın yok olmasına yol açmamamız için ahlaki bir zorunluluk veriyor.

Kozmoloji hakkında konferanslar verdiğimde, evrenimizin başka bir yerinde (Büyük Patlamamızdan bu yana 13,8 milyar yıl boyunca ışığın bize ulaştığı uzay bölgesi) zeki bir yaşam olduğunu düşünüyorlarsa, izleyicilere ellerini kaldırmalarını sık sık soruyorum. Şüphesiz, anaokullarından üniversite öğrencilerine kadar neredeyse herkes bunu yapıyor. Neden diye sorduğumda, alma eğiliminde olduğum temel cevap, Evrenimizin o kadar büyük olduğu ki, en azından istatistiksel olarak bir yerlerde yaşam olması gerektiğidir. Bu argümana daha yakından bakalım ve zayıflığını belirleyelim.

Her şey tek bir sayıya iniyor: bir medeniyet arasındaki tipik mesafe şekil 6.10 ve en yakın komşusu. Bu mesafe 20 milyar ışıkyılından çok daha büyükse, Evrenimizde (Büyük Patlamamızdan bu yana 13,8 milyar yıl boyunca ışığın bize ulaştığı uzayın parçası) yalnız kalmayı ve asla temas kurmamayı beklemeliyiz. uzaylılar. Peki bu mesafe için ne beklemeliyiz? Oldukça bilgisiziz. Bu, komşumuza olan mesafenin 1000 basketbol sahası içinde olduğu anlamına gelir. … Toplam sıfır sayısının makul olarak 21, 22, 23,…, 100, 101, 102 veya daha fazla olabileceği 000 metre - ancak muhtemelen 21'den küçük değil, çünkü henüz uzaylılara dair ikna edici kanıtlar görmedik (bkz.

şekil 6.11). En yakın komşu medeniyetimizin Evrenimizin içinde olması için, yarıçapı yaklaşık 10 olan 26 metre, sıfır sayısı 26'yı geçemez ve sıfırların sayısının 22-26 arasındaki dar aralıkta düşme olasılığı oldukça küçüktür. Bu yüzden Evrenimizde yalnız olduğumuzu düşünüyorum.



Şekil 6.11: Yalnız mıyız? Yaşamın ve zekânın nasıl evrimleştiğine dair büyük belirsizlikler, uzaydaki en yakın komşumuz uygarlığın yukarıdaki yatay eksen boyunca makul bir şekilde herhangi bir yerde olabileceğini düşündürüyor ve bu, bizim uçurumun kenarları arasındaki dar aralıkta olma ihtimalini düşük kılıyor.

Galaxy (yaklaşık 10^{21} metre uzaklıkta) ve Evrenimizin kenarı (yaklaşık 10^{26} metre uzaklıkta). Bu aralıktan çok daha yakın olsaydı, Galaksimizde muhtemelen fark edeceğimiz o kadar çok gelişmiş uygarlık olmalıydı ki bu da aslında Evrenimizde yalnız olduğumuzu gösteriyor.

Kitabımda bu argümanın ayrıntılı bir gerekçesini veriyorum *Matematiksel Evrenimiz,* bu yüzden burada tekrar etmeyeceğim, ancak bu komşu mesafesi hakkında bilgisiz olmamızın temel nedeni, belirli bir yerde zeki yaşamın ortaya çıkma olasılığı hakkında bilgisiz olmamızdır. Amerikalı gökbilimci Frank Drake'in belirttiği gibi, bu olasılık, orada yaşanabilir bir ortam olma olasılığı (örneğin uygun bir gezegen), orada yaşamın oluşma olasılığı ve bu yaşamın evrimleşme olasılığı ile çarpılarak hesaplanabilir. akıllı. Ben yüksek lisans öğrencisiyken, bu üç olasılık hakkında hiçbir fikrimiz yoktu. Geçtiğimiz yirmi yılda diğer yıldızların yörüngesinde dönen gezegenlerin dramatik keşiflerinden sonra, sadece kendi Galaksimizde milyarlarca yaşanabilir gezegenlerin bol olması muhtemel görünüyor. Bununla birlikte, yaşamın ve ardından zekanın evrimleşme olasılığı, son derece belirsizliğini koruyor: Bazı uzmanlar, birinin veya her ikisinin de kaçınılmaz olduğunu ve çoğu yaşanabilir gezegende meydana geldiğini düşünürken, diğerleri, bir veya daha fazla evrimsel darboğaz nedeniyle geçilmesi için vahşi bir şans gerektiren bir veya daha fazla darboğaz nedeniyle son derece nadir olduğunu düşünüyor. Önerilen bazı darboğazlar, en erken dönemde tavuk ve yumurta sorunlarını içerir.

kendi kendini yeniden üreten yaşamın aşamaları: örneğin, modern bir hücrenin bir ribozom inşa etmesi için, genetik kodumuzu okuyan ve proteinlerimizi oluşturan oldukça karmaşık moleküler makine için başka bir ribozoma ihtiyacı var ve çok açık değil.

ilk ribozom daha basit bir şeyden yavaş yavaş gelişebilirdi. [10] Önerilen diğer darboğazlar, daha yüksek zekanın geliştirilmesini içerir. Örneğin, dinozorlar Dünya'yı 100 milyon yıldan fazla bir süredir yönetiyor olsalar da, günümüz modern insanlarından bin kat daha uzun bir süredir, evrim onları kaçınılmaz olarak daha yüksek zekaya ve teleskoplar veya bilgisayarlar icat etmeye zorlamadı.

Bazı insanlar, evet, akıllı yaşam diyerek argümanıma karşı çıkıyor *abilir* çok nadir olabilir, ama aslında öyle değil - Galaksimiz, ana akım bilim adamlarının basitçe farketmediği zeki yaşamla doludur. Belki de UFO meraklılarının iddia ettiği gibi uzaylılar Dünya'yı çoktan ziyaret etmişlerdir. Belki de uzaylılar Dünya'yı ziyaret etmemişlerdir, ama oradalar ve kasıtlı olarak bizden saklanıyorlar (buna ABD'li gökbilimci John A. Ball tarafından "hayvanat bahçesi hipotezi" adı verildi ve bilim kurgu klasikleri gibi Olaf Stapledon *Star Maker).* Ya da belki de kasıtlı olarak saklanmadan oradalar: uzay yerleşimi veya fark ettiğimiz büyük mühendislik projeleriyle ilgilenmiyorlar.

Elbette, bu olasılıklar hakkında açık fikirli olmamız gerekiyor, ancak bunlardan herhangi biri için genel olarak kabul edilmiş bir kanıt olmadığından, alternatifi de ciddiye almalıyız: yalnızız. Dahası, ileri uygarlıkların çeşitliliğini, hepsinin fark edilmeden gitmelerine neden olan hedefleri paylaştığını varsayarak küçümsememeliyiz: Yukarıda gördük ki, bir uygarlık için kaynak edinme oldukça doğal bir hedeftir ve bizim fark etmemiz için, tek gereken *bir*

medeniyet alabildiği her şeyi ve dolayısıyla galaksimizi ve ötesini açıkça yerleştirmeye karar verir. Galaksimizde Dünya'dan milyarlarca yıl daha yaşlı olan ve hırslı sakinlerin Galaksiye yerleşmeleri için bolca zaman veren milyonlarca yaşanabilir Dünya benzeri gezegen olduğu gerçeğiyle karşı karşıya kaldığımızda, bu nedenle en bariz yorumu göz ardı edemeyiz: yaşamın kökeni rastgele bir tesadüf gerektirir, o kadar olası değildir ki, hepsi ıssızdır.

Hayat ise *değil* sonuçta nadir, yakında öğrenebiliriz. İddialı astronomik araştırmalar, yaşamın ürettiği oksijenin kanıtı için Dünya benzeri gezegenlerin atmosferlerini araştırıyor. Bu aramaya paralel olarak *hiç* hayat arayışı *akıllı* hayat kısa süre önce Rus hayırsever Yuri Milner'ın 100 milyon dolarlık projesi “Breakthrough Listen” ile canlandı.

Gelişmiş arama yaparken aşırı insan merkezli olmamak önemlidir.

yaşam: dünya dışı bir medeniyet keşfedersek, muhtemelen süper zekaya dönüşmüş olabilir. Martin Rees'in yakın tarihli bir makalesinde belirttiği gibi, "insan teknolojik medeniyetinin tarihi yüzyıllar içinde ölçülür - ve insanların inorganik zeka tarafından ele geçirilmesi veya aşılması, daha sonra varlığını sürdürecek ve gelişmeye devam edecek, yalnızca bir veya iki yüzyıl daha geçebilir. , milyarlarca yıl…. Onu kısa bir zaman diliminde 'yakalamamız' pek mümkün olmayacaktı.

organik biçim aldı. " [11] Jay Olson'ın yukarıda bahsedilen uzay yerleşim belgesinde şu sonuca katılıyorum: "Gelişmiş zekanın, evrenin kaynaklarını kullanarak mevcut dünyevi gezegenleri gelişmiş insan versiyonları ile basitçe doldurma olasılığını, teknolojinin ilerlemesinin olası olmayan bir son noktası olarak görüyoruz." Öyleyse uzaylıları hayal ettiğinizde, iki kollu ve iki bacağı olan küçük yeşil adamları düşünmeyin, bu bölümde daha önce keşfettiğimiz süper zeki uzay yolculuğu yaşamını düşünün.

Bilimdeki en büyüleyici sorulardan birine ışık tutan tüm dünya dışı yaşam arayışlarının güçlü bir destekçisi olsam da, gizlice hepsinin başarısız olacağını ve hiçbir şey bulamayacağını umuyorum! Galaksimizdeki yaşanabilir gezegenlerin bolluğu ile dünya dışı ziyaretçilerin yokluğu arasındaki bariz uyumsuzluk. *Fermi paradoksu,* iktisatçı Robin Hanson'un "Büyük Filtre" olarak adlandırdığı şeyin, cansız maddeden uzaya yerleşen yaşama gelişimsel yol boyunca bir yerde bir evrimsel / teknolojik engelin varlığını öne sürüyor. Başka bir yerde bağımsız olarak evrimleşmiş yaşamı keşfedersek, bu, ilkel yaşamın nadir olmadığını ve barikatın mevcut insan gelişim aşamasından sonra olduğunu gösterir - belki de uzayda yerleşim imkansız olduğundan veya neredeyse tüm gelişmiş medeniyetler kendilerinden önce kendi kendilerini yok ettiklerinden kozmik olabiliyor. Bu nedenle, dünya dışı yaşam arayışlarının hiçbir şey bulamadığı için parmaklarımı uzatıyorum: bu, akıllı yaşamın evrimleşmesinin nadir olduğu ama biz insanların şanslı olduğu senaryo ile tutarlı, böylece arkamızda engel var ve olağanüstü gelecek potansiyeline sahibiz.

Görünüm

Şimdiye kadar, bu kitabı milyarlarca yıl önceki mütevazı başlangıcından milyarlarca yıl sonraki olası büyük geleceklere kadar Evrenimizdeki yaşamın tarihini keşfetmek için harcadık. Mevcut AI geliştirmemiz sonunda bir istihbarat patlamasını ve optimize edilmiş uzay yerleşimini tetiklerse, bu gerçekten kozmik anlamda bir patlama olacaktır: Kayıtsız cansız bir kozmosta neredeyse ihmal edilebilir derecede küçük bir tedirginlik olarak milyarlarca yıl geçirdikten sonra, hayat aniden kozmik arenaya patlar. ışık hızına yakın genişleyen, asla yavaşlamayan ve yoluna çıkan her şeyi yaşam kıvılcımı ile ateşleyen küresel bir patlama dalgası olarak.

Kozmik geleceğimizdeki yaşamın önemine ilişkin bu tür iyimser görüşler, bu kitapta karşılaştığımız birçok düşünür tarafından güzel bir şekilde ifade edilmiştir. Bilim kurgu yazarları genellikle gerçekçi olmayan romantik hayalperestler olarak reddedildikleri için, uzay yerleşimi hakkındaki çoğu bilimkurgu ve bilimsel yazının şimdi de görünmesini ironik buluyorum. *karamsar* süper zekanın ışığında. Örneğin, insanlar ve diğer zeki varlıklar dijital biçimde aktarıldıktan sonra galaksiler arası seyahatin çok daha kolay hale geldiğini gördük, bu da bizi potansiyel olarak yalnızca Güneş Sistemimizde veya Samanyolu Galaksisinde değil, aynı zamanda kozmosta da kendi kaderimizin efendisi yapıyor.

Yukarıda, Evrenimizdeki tek yüksek teknoloji uygarlığı olduğumuzun gerçek olasılığını düşündük. Bu bölümün geri kalanını bu senaryoyu ve bunun gerektirdiği büyük ahlaki sorumluluğu araştırarak geçirelim. Bu, 13,8 milyar yıldan sonra, Evrenimizdeki yaşamın, kozmosta gelişmekle nesli tükenmek arasında bir seçimle karşı karşıya olan yolda bir çatala ulaştığı anlamına gelir. Teknolojimizi geliştirmeye devam etmezsek, soru insanlığın yok olup olmayacağı değil, *Nasıl.* Bizi ilk önce ne yakalayacak - bir asteroit, bir süper yanardağ, yaşlanan Güneş'in yakıcı ısısı veya başka bir felaket (bkz. şekil 5.1)? Bir kez gittiğimizde, Freeman Dyson tarafından tahmin edilen kozmik drama seyirci olmadan devam edecek: bir kozmokalip olmadan, yıldızlar yanar, galaksiler kaybolur ve kara delikler buharlaşır, her biri hayatını bir milyondan fazla kez salınan büyük bir patlamayla sonlandırır. Tsar Bomba olarak enerji, şimdiye kadar yapılmış en güçlü hidrojen bombası. Freeman'ın dediği gibi: "Soğuk genişleyen evren, çok uzun bir süre ara sıra havai fişeklerle aydınlatılacak." Ne yazık ki, bu havai fişek gösterisinin tadını çıkaracak kimse olmadığı için anlamsız bir israf olacak.

Teknoloji olmadan, insan neslinin tükenmesi, on milyarlarca yıllık kozmik bağlamda çok yakındır ve Evrenimizdeki tüm yaşam dramını, hiç kimsenin deneyimlemediği neredeyse sonsuz bir anlamsızlık içinde kısa ve geçici bir güzellik, tutku ve anlam parıltısı haline getirir. Bu ne kadar boşa bir fırsat olurdu! Teknolojiden kaçınmak yerine, onu benimsemeyi seçersek, o zaman yükseliriz: Hem yaşamın hayatta kalma ve gelişme potansiyeli hem de yaşamın daha da erken tükenmesi için potansiyel kazanırız, kötü planlama nedeniyle kendi kendini yok eder (bkz. şekil 5.1). Benim oyum teknolojiyi kucaklamak ve inşa ettiğimiz şeye körü körüne inançla değil, dikkatli, öngörü ve dikkatli planlama ile ilerlemek.

13.8 milyar yıllık kozmik tarihten sonra, kendimizi nefes kesici güzellikte bir Evrende buluyoruz, bu Evren bizim aracılığımızla insanlar canlanıyor ve kendisinin farkına varmaya başlıyor. Evrenimizdeki yaşamın gelecekteki potansiyelinin atalarımızın en çılgın hayallerinden daha büyük olduğunu gördük, zeki yaşamın kalıcı olarak yok olması için eşit derecede gerçek bir potansiyel tarafından yumuşatıldı. Evrenimizdeki yaşam potansiyelini gerçekleştirecek mi yoksa israf mı edecek? Bu büyük ölçüde bugün yaşayan biz insanların yaşamımız boyunca ne yaptığına bağlıdır ve doğru seçimleri yaparsak yaşamın geleceğini gerçekten harika hale getirebileceğimiz konusunda iyimserim. Ne istemeliyiz ve bu hedeflere nasıl ulaşabiliriz? Kitabın geri kalanını ilgili en zor zorluklardan bazılarını ve bunlar hakkında neler yapabileceğimizi araştırarak geçirelim.

ALT ÇİZGİ:

- Milyarlarca yıllık kozmik zaman ölçekleriyle karşılaştırıldığında, bir istihbarat patlaması, teknolojinin yalnızca fizik yasalarıyla sınırlı bir düzeyde hızla plato yaptığı ani bir olaydır.

- Bu teknolojik plato, günümüz teknolojisinden çok daha yüksektir ve belirli bir miktar maddenin yaklaşık on milyar kat daha fazla enerji üretmesine (sfalerin veya kara deliklerin kullanılmasıyla), 12–18 büyüklüğünde daha fazla bilgi depolamasına veya 31–41 büyüklük mertebesini hesaplamasına izin verir. daha hızlı - ya da istenen herhangi bir başka madde biçimine dönüştürülmek.

- Süper zeki yaşam, mevcut kaynaklarını böylesine çarpıcı biçimde daha verimli kullanmakla kalmayacak, aynı zamanda ışık hızına yakın kozmik yerleşim yoluyla daha fazla kaynak elde ederek bugünün biyosferini yaklaşık 32 mertebesinde büyütebilecektir.

- Karanlık enerji, süper zeki yaşamın kozmik genişlemesini sınırlar ve ayrıca onu uzaktaki genişleyen ölüm balonlarından veya düşman uygarlıklardan korur. Karanlık enerjinin kozmik medeniyetleri parçalaması tehdidi, eğer mümkünse solucan deliği inşaatı da dahil olmak üzere devasa kozmik mühendislik projelerini motive ediyor.

- Kozmik mesafelerde paylaşılan veya ticareti yapılan ana meta muhtemelen bilgi olacaktır.

- Solucan delikleri dışında, iletişimdeki ışık hızı sınırı, kozmik bir uygarlık genelinde koordinasyon ve kontrol için ciddi zorluklar ortaya çıkarır. Uzaktaki bir merkez merkez, süper zeki "düğümlerini", örneğin kurallara uyulmadığı sürece bir süpernova veya kuasar başlatarak düğümü yok etmeye programlanmış yerel bir koruma YZ'si konuşlandırarak, ödüller veya tehditler yoluyla işbirliği yapmaya teşvik edebilir.

- Genişleyen iki medeniyetin çarpışması, asimilasyon, işbirliği veya savaşla sonuçlanabilir, burada ikincisi tartışmalı olarak bugünün medeniyetleri arasında olduğundan daha az olasıdır.

- Aksine yaygın inanışa rağmen, gözlemlenebilir Evrenimizi gelecekte canlandırabilecek tek yaşam formunun biz olduğumuz oldukça makul.

- Teknolojimizi geliştirmezsek, soru insanlığın yok olup olmayacağı değil, sadece nasıl: Bir asteroid, bir süper yanardağ, yaşlanan Güneş'in yakıcı ısısı veya başka bir felaket bizi ilk önce kurtaracak mı?

- Teknolojimizi tuzaklardan kaçınmak için yeterli özen, öngörü ve planlama ile geliştirmeye devam edersek, yaşamın Dünya'da ve atalarımızın en çılgın hayallerinin ötesinde milyarlarca yıl boyunca gelişme potansiyeli vardır.

* 1 Enerji sektöründe çalışıyorsanız, bunun yerine verimliliği, salınan enerjinin yararlı bir biçimde olan kısmı olarak tanımlamaya alışkın olabilirsiniz.

* 2 Yakındaki evrende uygun bir doğa-yapımı kara delik bulunamazsa, yeterince küçük bir alana çok miktarda madde koyarak yeni bir tane yaratılabilir.

* 3 Bu biraz fazla basitleştirmedir, çünkü Hawking radyasyonu aynı zamanda yararlı iş çıkarmanın zor olduğu bazı parçacıkları da içerir. Büyük kara delikler yalnızca% 90 verimlidir, çünkü enerjinin yaklaşık% 10'u graviton biçiminde yayılır: yararlı iş çıkarmaları bir yana, tespit edilmesi neredeyse imkansız olan son derece utangaç parçacıklar. Kara delik buharlaşmaya ve küçülmeye devam ettikçe, Hawking radyasyonu nötrinoları ve diğer büyük parçacıkları içerdiği için verimlilik daha da düşer.

* 4 Dışarıdaki Douglas Adams hayranları için, bunun yaşam, evren ve her şey sorusuna cevap veren zarif bir soru olduğuna dikkat edin. Daha doğrusu, verimlilik $1 - 1 / \sqrt{3} \approx$% 42'dir.

* 5 Kara deliği, etrafına aynı yönde yavaşça dönen bir gaz bulutu yerleştirerek beslerseniz, bu gaz çekilip yenildikçe daha da hızlı dönerek karadeliğin dönüşünü hızlandırır, tıpkı bir figür patencisinin daha hızlı dönmesi gibi. kollarını çekerek. Bu, deliğin maksimum şekilde dönmesini sağlayarak, önce gaz enerjisinin% 42'sini ve sonra geri kalanının% 29'unu toplam verim için% 42 + (% 1-% 42) ×% 29 ≈% 59 çıkarmanıza olanak tanır.

* 6 Elektromanyetik ve zayıf kuvvetleri yeniden birleştirmek için yeterince ısınması gerekiyor; bu, parçacıklar bir parçacık çarpıştırıcısında 200 milyar volt hızlandıklarında olduğu kadar hızlı hareket ettiğinde meydana geliyor.

* 7 Yukarıda sadece atomlardan oluşan maddeyi tartıştık. Yaklaşık altı kat daha fazla karanlık madde var, ancak çok zor ve yakalanması zor, rutin olarak doğrudan Dünya'nın içinden ve diğer taraftan uçuyor, bu nedenle gelecekteki yaşamın onu yakalayıp kullanmasının mümkün olup olmadığı görülmeye devam ediyor.

* 8 Kozmik matematik son derece basit çıkıyor: eğer medeniyet ışık hızında değil, genişleyen uzayda genişlerse *c* ama daha yavaş bir hızda *v,* yerleştirilen galaksi sayısı

bir faktörle azaltılmış ($(v / c)^3$. Bu, yavaş ilerleyen medeniyetlerin ağır şekilde cezalandırıldığı anlamına gelir; 10 kat daha yavaş genişleyen bir uygarlık, sonuçta 1000 kat daha az galaksiye yerleşir.

* 9 Bununla birlikte, John Gribbin 2011 kitabında benzer bir sonuca varıyor *Evrende yalnız.* Bu soruya ilişkin ilgi çekici perspektiflerden oluşan bir yelpaze için Paul Davies'in 2011 kitabını da öneriyorum. *Ürkütücü Sessizlik.*

Bölüm 7

Hedefler

İnsan varoluşunun gizemi sadece hayatta kalmakta değil, yaşamak için bir şeyler bulmakta yatmaktadır.

Fyodor Dostoyevski, *Karamazov Kardeşler*

Hayat bir yolculuktur, varış noktası değil.

Ralph Waldo Emerson

En çetrefilli yapay zeka tartışmalarının ne hakkında olduğunu tek bir kelimeyle özetlemek zorunda kalsaydım, bu "hedefler" olurdu: YZ hedefleri vermeli miyiz ve öyleyse, kimin hedefleri? AI hedeflerini nasıl verebiliriz? AI daha akıllı hale gelse bile bu hedeflerin korunmasını sağlayabilir miyiz? Bizden daha akıllı bir yapay zekanın hedeflerini değiştirebilir miyiz? Nihai hedeflerimiz nelerdir? Bu sorular sadece zor değil, aynı zamanda yaşamın geleceği için de çok önemlidir: Ne istediğimizi bilmiyorsak, onu elde etme olasılığımız azalır ve kontrolü hedeflerimizi paylaşmayan makinelere bırakırsak, o zaman istemediğimiz şeyi elde etmemiz muhtemeldir.

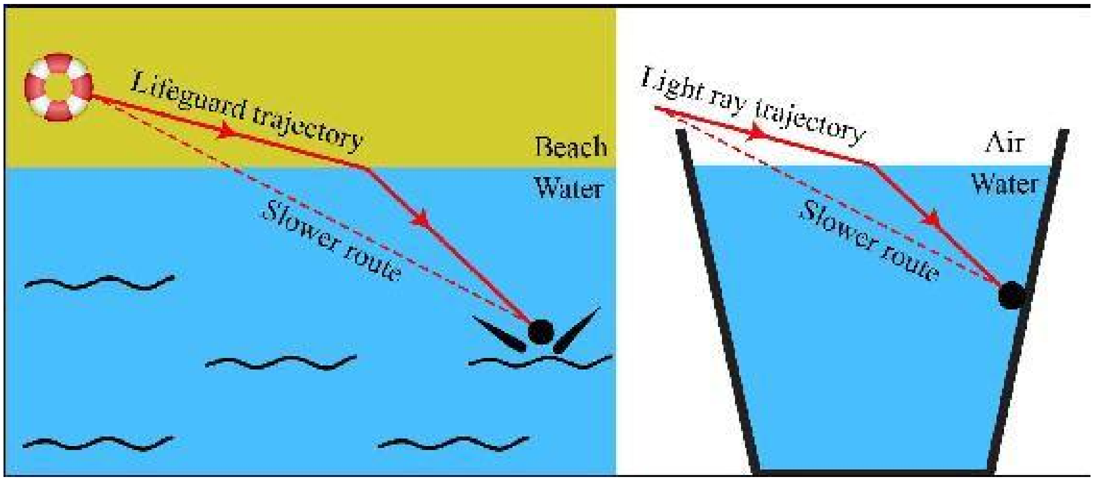
Fizik: Hedeflerin Kökeni

Bu sorulara ışık tutmak için önce hedeflerin nihai kökenini inceleyelim. Dünyada etrafımıza baktığımızda bazı süreçler bize şu şekilde çarpıyor: *hedef odaklı*
diğerleri yapmazken. Örneğin, oyunu kazandıran atış için bir futbol topunun atılma sürecini düşünün. Topun davranışı hedefe yönelik görünmüyor ve en ekonomik olarak vuruşa bir tepki olarak Newton'un hareket yasaları ile açıklanıyor. Oyuncunun davranışı ise en ekonomik olarak mekanik olarak değil, birbirini iten atomlar açısından değil, *hedef* takımının puanını en üst düzeye çıkarmak. Bu tür hedef odaklı davranış, yalnızca görünüşte amaçsız olarak etrafta zıplayan bir grup parçacığın oluşturduğu erken Evrenimizin fiziğinden nasıl ortaya çıktı?

Şaşırtıcı bir şekilde, hedefe yönelik davranışın nihai kökleri fizik kanunlarında bulunabilir ve kendilerini hayatı içermeyen basit süreçlerde bile gösterebilir. Bir cankurtaran, bir yüzücüyü kurtarırsa, şekil 7.1 , düz bir çizgide gitmemesini, sudan daha hızlı gidebileceği sahil boyunca biraz daha ileriye koşmasını, böylece suya girdiğinde hafifçe dönmesini bekliyoruz. Onun yörünge seçimini doğal olarak hedef odaklı olarak yorumluyoruz, çünkü tüm olası yörüngelerden, onu yüzücüye olabildiğince hızlı ulaştıracak en uygun yolu kasıtlı olarak seçiyor. Yine de basit bir ışık huzmesi suya girdiğinde benzer şekilde bükülür (bkz. şekil 7.1), ayrıca varış noktasına seyahat süresini de en aza indirir! Bu nasıl olabilir?

Bu fizikte şu şekilde bilinir *Fermat prensibi,* 1662'de eklemlenmiştir ve ışık ışınlarının davranışını tahmin etmenin alternatif bir yolunu sağlar. Dikkat çekici bir şekilde, fizikçiler o zamandan beri bunu keşfettiler *herşey* Klasik fizik yasaları benzer bir şekilde matematiksel olarak yeniden formüle edilebilir: doğanın bir şeyi yapmayı seçebileceği her yolun dışında, tipik olarak bazı miktarı en aza indirmeye veya en üst düzeye çıkarmaya indirgeyen en uygun yolu tercih eder. Her bir fiziksel yasayı tanımlamanın matematiksel olarak eşdeğer iki yolu vardır: ya geçmişin geleceğe neden olması ya da bir şeyi optimize eden doğa olarak. İkinci yol genellikle giriş seviyesi fizik derslerinde öğretilmese de matematik daha zor olduğu için, daha zarif ve derin olduğunu hissediyorum. Bir kişi bir şeyi optimize etmeye çalışıyorsa (örneğin puanını, servetini veya mutluluğunu) doğal olarak açıklayacağız

hedef odaklı olarak arayışları. Dolayısıyla, eğer doğanın kendisi bir şeyi optimize etmeye çalışıyorsa, o zaman hedefe yönelik davranışın ortaya çıkmasına şaşmamalı: bu, en başından itibaren, fizik kanunlarına dahil edilmiştir.



Şekil 7.1: Bir yüzücüyü olabildiğince hızlı kurtarmak için, cankurtaran düz bir çizgide (kesikli) gitmez, ancak sahil boyunca sudakinden daha hızlı gidebileceği biraz daha uzağa gider. Bir ışık ışını, hedefine olabildiğince hızlı ulaşmak için suya girerken benzer şekilde bükülür.

Doğanın maksimize etmeye çalıştığı ünlü bir nicelik, *entropi,* Bu gevşek bir şekilde işlerin ne kadar dağınık olduğunu ölçer. Termodinamiğin ikinci yasası, entropinin mümkün olan maksimum değerine ulaşana kadar artma eğiliminde olduğunu belirtir. Şimdilik yerçekiminin etkilerini göz ardı ederek, bu maksimum dağınık son duruma

ısı ölümü, ve karmaşıklık, yaşam ve değişim olmaksızın, sıkıcı mükemmel bir tekdüzelik içinde yayılan her şeye karşılık gelir. Örneğin sıcak kahveye soğuk süt döktüğünüzde, içeceğinizin geri dönüşü olmayan bir şekilde kendi kişisel ısı ölümü hedefine doğru ilerliyor gibi görünüyor ve çok geçmeden, hepsi tek tip ılık bir karışım. Canlı bir organizma ölürse, entropisi de yükselmeye başlar ve çok geçmeden parçacıklarının dizilişi çok daha az organize olma eğilimindedir.

Doğanın entropiyi artırmaya yönelik görünürdeki hedefi, zamanın neden tercih edilen bir yöne sahip olduğunu açıklamaya yardımcı olur ve filmlerin geriye doğru oynatıldığında gerçekçi görünmemesine neden olur: bir bardak şarap düşürürseniz, yere çarpıp küresel dağınıklığı (entropi) artırmasını beklersiniz. Sonra gördüysen *parçalanmamış* ve elinize sağlam bir şekilde geri uçarsanız (entropiyi azaltır), muhtemelen çok fazla bardak içtiğinizi düşünerek onu içmezsiniz.

Sıcaktan ölüme doğru amansız ilerlememizi ilk öğrendiğimde, bunu oldukça iç karartıcı buldum ve yalnız değildim: termodinamiğin öncüsü Lord Kelvin 1841'de şöyle yazmıştı: “Sonuç kaçınılmaz olarak evrensel bir dinlenme hali olacaktır.

ve ölüm ”ve doğanın uzun vadeli amacının ölüm ve yıkımı maksimize etmek olduğu fikrinde teselli bulmak zor. Ancak, daha yeni keşifler, işlerin o kadar da kötü olmadığını gösterdi. Her şeyden önce, yerçekimi diğer tüm kuvvetlerden farklı davranır ve Evrenimizi daha tekdüze ve sıkıcı değil, daha yığılmış ve ilginç hale getirmeye çalışır. Bu nedenle yerçekimi, neredeyse mükemmel bir şekilde tekdüze olan sıkıcı erken Evrenimizi, galaksiler, yıldızlar ve gezegenlerle dolu, günümüzün yığınlı ve güzel bir şekilde karmaşık kozmosuna dönüştürdü. Yerçekimi sayesinde, artık sıcak ve soğuğu birleştirerek yaşamın gelişmesine izin veren geniş bir sıcaklık aralığı var: 6.000 ° C (10.000 ° F) güneş ısısını emen, rahat ve sıcak bir gezegende yaşıyoruz ve atık ısıyı soğuk uzaya yayarak serinliyoruz. sıcaklık mutlak sıfırın sadece 3 ° C (5 ° F) üzerindedir.

İkincisi, MIT'deki meslektaşım Jeremy England ve diğerlerinin son çalışmaları, termodinamiğin doğaya da katkı sağladığını gösteren daha iyi haberler getirdi.
ısı ölümünden daha ilham verici bir hedef. [1] Bu hedefe inek adı verilir
dağıtım odaklı adaptasyon, Bu, temel olarak, rastgele parçacık gruplarının, enerjiyi çevrelerinden olabildiğince verimli bir şekilde çıkarmak için kendilerini organize etmeye çalıştıkları anlamına gelir ("yayılma", genellikle işlemde faydalı işler yaparken, genellikle yararlı enerjiyi ısıya dönüştürerek entropinin artmasına neden olmak anlamına gelir.). Örneğin, güneş ışığına maruz kalan bir grup molekül, zamanla güneş ışığını daha iyi ve daha iyi absorbe etme eğiliminde olacaktır. Başka bir deyişle, doğa giderek daha karmaşık ve gerçeğe yakın olan kendi kendini organize eden sistemler üretmeye yönelik yerleşik bir hedefe sahip gibi görünmektedir ve bu amaç, fiziğin kanunlarına dahil edilmiştir.

Yaşama yönelik bu kozmik dürtü ile ısı ölümüne yönelik kozmik dürtüyü nasıl uzlaştırabiliriz? Cevap, 1944 tarihli ünlü kitapta bulunabilir. *Hayat nedir?* kuantum mekaniğinin kurucularından Erwin Schrödinger tarafından. Schrödinger, yaşayan bir sistemin ayırt edici özelliklerinden birinin, entropisini etrafındaki entropiyi artırarak sürdürmesi veya azaltması olduğuna dikkat çekti. Başka bir deyişle, termodinamiğin ikinci yasasının bir yaşam boşluğu vardır: Toplam entropinin artması gerekmesine rağmen, başka yerlerde daha da arttığı sürece bazı yerlerde azalmasına izin verilir. Dolayısıyla hayat, çevresini daha da karmaşık hale getirerek karmaşıklığını sürdürür veya artırır.

Biyoloji: Hedeflerin Evrimi

Hedefe yönelik davranışın kökeninin nasıl mümkün olduğunca verimli bir şekilde çevrelerinden enerji elde etmek için kendilerini düzenleme amacıyla parçacıklara bahşeden fizik yasalarına kadar nasıl izlenebileceğini gördük. Bir parçacık düzenlemesinin bu hedefi daha da ilerletmesinin harika bir yolu, daha fazla enerji soğurucu üretmek için kendi kopyalarını yapmaktır. Bu tür ortaya çıkan kendi kendini kopyalamanın bilinen birçok örneği vardır: örneğin, türbülanslı sıvılardaki girdaplar kendi kopyalarını oluşturabilir ve mikro küre kümeleri, yakındaki küreleri aynı kümeler oluşturacak şekilde koaksiyel olabilir. Bir noktada, belirli bir parçacık düzenlemesi kendisini kopyalamada o kadar başarılı oldu ki, bunu çevresinden enerji ve hammadde çıkararak neredeyse sonsuza kadar yapabilirdi. Böyle bir parçacık düzenlemesi diyoruz *hayat.* Dünyada yaşamın nasıl ortaya çıktığı hakkında hala çok az şey biliyoruz, ancak ilkel yaşam formlarının yaklaşık 4 milyar yıl önce burada olduğunu biliyoruz.

Bir yaşam formu kendini kopyalarsa ve kopyalar da aynısını yaparsa, toplam sayı, nüfus boyutu kaynak sınırlamalarına veya diğer sorunlara karşı çıkana kadar düzenli aralıklarla ikiye katlanmaya devam edecektir. Tekrar tekrar ikiye katlama kısa sürede çok büyük rakamlar üretir: Bir ile başlar ve sadece üç yüz kez iki katına çıkarırsanız, Evrenimizdeki parçacık sayısını aşan bir miktar elde edersiniz. Bu, ilk ilkel yaşam formunun ortaya çıkmasından kısa bir süre sonra, büyük miktarlarda maddenin canlandığı anlamına gelir. Bazen kopyalama mükemmel değildi, bu yüzden kısa sürede kendilerini kopyalamaya çalışan, aynı sınırlı kaynaklar için rekabet eden birçok farklı yaşam formu ortaya çıktı. Darwinci evrim başlamıştı.

Yaşamın başladığı sıralarda Dünya'yı sessizce gözlemliyor olsaydınız, hedefe yönelik davranışta dramatik bir değişiklik fark ederdiniz. Daha önce, parçacıklar ortalama dağınıklığı çeşitli şekillerde artırmaya çalışıyormuş gibi görünürken, bu yeni her yerde bulunan kendi kendini kopyalayan modellerin farklı bir amacı varmış gibi görünüyordu: *çoğaltma.* Charles Darwin nedenini zarif bir şekilde açıkladı: En verimli fotokopi makineleri diğerlerini geride bırakıp onlara hükmettiği için, çok geçmeden baktığınız herhangi bir rastgele yaşam formu, çoğaltma hedefi için oldukça optimize edilmiş olacaktır.

Fizik yasaları aynı kaldığında, amaç dağılmadan kopyalamaya nasıl değişebilirdi? Cevap şu ki, temel amaç (dağılma)

yapmadı değişti, ancak farklı bir *enstrümantal hedef* yani, temel hedefe ulaşılmasına yardımcı olan bir alt hedef. Örneğin yemek yemeyi ele alalım. Evrimin tek temel amacının çiğneme değil, kopyalama olduğunu bilsek de, hepimizin açlık arzumuzu tatmin etme hedefi varmış gibi görünüyor. Bunun nedeni, yemek yemenin çoğalmaya yardımcı olmasıdır: açlıktan ölmek çocuk sahibi olmanın önüne geçer. Aynı şekilde, çoğaltma dağılmaya yardımcı olur, çünkü yaşamla iç içe olan bir gezegen enerji yaymada daha etkilidir. Yani bir anlamda evrenimiz, sıcak ölüme daha hızlı yaklaşmasına yardımcı olmak için yaşamı icat etti. Mutfağınızın zeminine şeker dökerseniz, prensipte faydalı kimyasal enerjisini yıllarca koruyabilir, ancak karıncalar ortaya çıkarsa, bu enerjiyi hiçbir zaman yok ederler. Benzer şekilde,

Bugünün gelişmiş Dünya sakinleri arasında, bu araçsal hedefler kendi başlarına bir hayata kavuşmuş gibi görünüyor: evrim onları yalnızca kopyalama hedefi için optimize etse de, çoğu zamanlarının çoğunu yavru üretmek için değil, uyku, yiyecek peşinde koşma gibi faaliyetlere harcıyor. , evler inşa etmek, egemenlik kurmak ve savaşmak ya da başkalarına yardım etmek - hatta bazen
azaltır çoğaltma. Evrimsel psikoloji, ekonomi ve yapay zeka alanındaki araştırmalar, nedenini zarif bir şekilde açıkladı. Bazı iktisatçılar, insanları rasyonel failler olarak modelliyorlardı, hedeflerine ulaşmak için her zaman en uygun eylemi seçen idealleştirilmiş karar vericiler, ama bu kesinlikle gerçekçi değil. Uygulamada, bu temsilciler, Nobel ödüllü ve yapay zeka öncüsü Herbert Simon'un "sınırlı rasyonalite" olarak adlandırdıkları şeye sahipler çünkü sınırlı kaynakları var: Kararlarının rasyonelliği, mevcut bilgileriyle, düşünme için mevcut zamanlarıyla ve düşünebilecekleri mevcut donanımlarıyla sınırlıdır. . Bu, Darwinci evrimin bir organizmayı bir hedefe ulaşmak için optimize ederken yapabileceği en iyi şeyin, ajanın tipik olarak kendini bulduğu kısıtlı bağlamda makul derecede iyi çalışan yaklaşık bir algoritma uygulamak olduğu anlamına gelir. Evrim, çoğaltma optimizasyonunu tam olarak şu şekilde uygulamıştır: Her durumda, hangi eylemin bir organizmanın başarılı yavru sayısını en üst düzeye çıkaracağını sormak yerine, bir sezgisel korsanlık hilesi uygular: genellikle iyi çalışan pratik kurallar. Çoğu hayvan için bunlar arasında cinsel dürtü, susadığında içki içmek, açken yemek yemek ve tadı kötü ya da inciten şeylerden kaçınmak yer alır.

Bu temel kurallar bazen başa çıkmak için tasarlanmadıkları durumlarda, örneğin farelerin lezzetli tada sahip fare zehiri yediğinde, güvelerin baştan çıkarıcı dişi kokuları tarafından tutkal tuzaklarına çekildiği ve böcekler uçtuğunda kötü bir şekilde başarısız olur.

mum alevlerinin içine. * [1] Günümüz insan toplumu, evrimin temel kurallarımızı optimize ettiği çevreden çok farklı olduğundan, davranışımızın genellikle bebek yapmayı en üst düzeye çıkarmakta başarısız olduğunu görünce şaşırmamalıyız. Örneğin, açlıktan ölmeme alt hedefi kısmen kalorili yiyecekleri tüketme arzusu olarak uygulanmakta ve günümüzün obezite salgınını ve flört zorluklarını tetiklemektedir. Üreme alt hedefi, daha az çabayla daha fazla bebek üretebilse de, sperm / yumurta donörü olma arzusundan çok seks arzusu olarak uygulandı.

Psikoloji: Hedeflerin Peşinde ve İsyan

Özetle, canlı bir organizma, tek bir hedef peşinde koşmayan, bunun yerine neyin peşinden gideceği ve neyin kaçınılacağı konusunda genel kuralları izleyen sınırlı bir akılcılığın aracıdır. İnsan zihnimiz bu evrimleşmiş temel kuralları şu şekilde algılar: *duygular*

bu genellikle (ve çoğu kez farkında olmadan) nihai çoğaltma hedefine doğru karar vermemize rehberlik eder. Açlık ve susuzluk duyguları bizi açlık ve susuzluktan korur, acı duyguları bizi bedenlerimize zarar vermekten korur, şehvet duyguları bizi doğurur, sevgi ve şefkat duyguları bizi genlerimizin diğer taşıyıcılarına ve onlara yardım edenlere yardım eder vb. . Bu duyguların rehberliğinde beyinlerimiz, her seçimi kaç tane torun üreteceğimize dair nihai sonuçlarının sıkıcı bir analizine tabi tutmak zorunda kalmadan hızlı ve verimli bir şekilde ne yapacaklarına karar verebilir. Duygular ve fizyolojik kökenleri hakkında yakından ilişkili bakış açıları için,

William James ve António Damásio'nun yazılarını öneriyoruz. [2]

Duygularımızın ara sıra işe yaradığına dikkat etmek önemlidir. *karşısında* bebek yapmak, bu ille de kazara ya da kandırıldığımız için değil: beynimiz genlerimize ve onların çoğaltma hedeflerine oldukça bilinçli bir şekilde isyan edebilir, örneğin doğum kontrol hapı kullanmayı seçerek! Beynin genlerine başkaldırmasının daha aşırı örnekleri arasında intihar etmeyi veya bir rahip, keşiş veya rahibe olmak için yaşamını bekârlıkla geçirmeyi seçmeyi içerir.

Neden bazen genlerimize ve onların çoğaltma hedeflerine isyan etmeyi seçiyoruz? İsyan ediyoruz çünkü tasarım gereği, sınırlı akılcılığın temsilcileri olarak, sadece duygularımıza sadıkız. Beynimiz sadece genlerimizi kopyalamaya yardımcı olmak için evrimleşmiş olsa da, genlerle ilgili hiçbir duyguya sahip olmadığımız için beynimiz bu hedefle daha az ilgilenemezdi - aslında, insanlık tarihinin çoğunda atalarımız bunların onların olduğunu bile bilmiyorlardı. *vardı* genler. Dahası, beyinlerimiz genlerimizden çok daha akıllıdır ve artık genlerimizin amacını (çoğaltma) anladığımıza göre, bunu oldukça sıradan ve görmezden gelmenin kolay olduğunu düşünüyoruz. İnsanlar, genlerinin neden kendilerine şehvet hissettirdiğini anlayabilir, ancak on beş çocuk yetiştirmek için çok az istek duyabilirler ve bu nedenle, yakınlığın duygusal ödüllerini doğum kontrolüyle birleştirerek genetik programlarını kırmayı seçebilirler. Genlerinin neden şekerleri canlandırdığını ancak kilo alma arzusunun az olduğunu anlayabilirler ve bu nedenle, tatlı bir içeceğin duygusal ödüllerini

sıfır kalorili yapay tatlandırıcılar.

İnsanlar eroine bağımlı hale geldiklerinde olduğu gibi, bu tür ödül-mekanizma saldırıları bazen ters gitse de, insan gen havuzumuz kurnaz ve asi beyinlerimize rağmen şu ana kadar gayet iyi bir şekilde hayatta kaldı. Bununla birlikte, nihai otoritenin artık genlerimiz değil, duygularımız olduğunu hatırlamak önemlidir. Bu, insan davranışının kesinlikle türümüzün hayatta kalması için optimize edilmediği anlamına gelir. Aslında, duygularımız her durumda uygun olmayan sadece pratik kuralları uyguladığından, insan davranışının kesinlikle iyi tanımlanmış tek bir amacı yoktur.

Mühendislik: Dış Kaynak Kullanımı Hedefleri

Makinelerin hedefleri olabilir mi? Bu basit soru büyük tartışmalara yol açtı, çünkü farklı insanlar bunu farklı anlamlar olarak kabul ediyor, genellikle makinelerin bilinçli olup olamayacağı ve duyguları olup olmadığı gibi çetrefilli konularla ilgili. Ancak daha pratik isek ve soruyu basitçe "Makineler hedefe yönelik davranış sergileyebilir mi?" Anlamında alırsak, cevap açıktır: "Elbette yapabilirler, çünkü biz onları bu şekilde tasarlayabiliriz!" Fare kapanları, bulaşıkları temizlemek amacıyla bulaşık makineleri ve zamanı tutma hedefiyle saat yakalama hedefine sahip fare kapanları tasarlıyoruz. Bir makineyle yüzleştiğinizde, onun hedefe yönelik davranış sergilediği ampirik gerçeği, genellikle tek umursadığınız şeydir: eğer ısı arayan bir füze tarafından kovalanırsanız, bilinçli olup olmadığı veya hisleri olup olmadığı umurunuzda değildir!

Şimdiye kadar inşa ettiğimiz şeylerin çoğu sadece hedef odaklı sergiler *tasarım* hedef odaklı değil *davranış:* bir otoyol davranmaz; sadece orada oturuyor. Bununla birlikte, varlığının en ekonomik açıklaması, bir hedefe ulaşmak için tasarlanmış olmasıdır, bu nedenle bu tür pasif teknoloji bile Evrenimizi daha hedef odaklı hale getiriyor. *Teleoloji* şeylerin nedenlerinden ziyade amaçları açısından açıklamasıdır, bu nedenle bu bölümün ilk bölümünü Evrenimizin giderek daha teleolojik hale geldiğini söyleyerek özetleyebiliriz.

Sadece değil *Yapabilmek* cansız maddenin en azından bu zayıf anlamda hedefleri vardır, ancak giderek *yapar.* Gezegenimiz oluştuğundan beri Dünya'nın atomlarını gözlemliyor olsaydınız, hedefe yönelik davranışın üç aşamasını fark etmişsinizdir:

1. Bütün mesele dağılmaya odaklanmış gibiydi (entropi artışı).
2. Meselenin bir kısmı canlandı ve bunun yerine bunun kopyalarına ve alt hedeflerine odaklandı.
3. Hedeflerine ulaşmalarına yardımcı olmak için hızla büyüyen bir madde fraksiyonu canlı organizmalar tarafından yeniden düzenlendi.

Tablo 7.1 insanlığın fizikten nasıl baskın hale geldiğini gösterir

bakış açısı: şu anda sadece inekler hariç diğer tüm memelilerden daha fazla madde içermiyoruz (sığır eti ve süt ürünlerini tüketme hedeflerimize hizmet ettikleri için çok sayıdadırlar), aynı zamanda makinelerimizdeki, yollarımızdaki, binalarımızdaki ve diğer mühendislik projelerimizdeki konu Yakında Dünya'daki tüm canlı maddeyi ele geçirmenin yolu. Diğer bir deyişle, bir istihbarat patlaması olmasa bile, Dünya'daki hedefe yönelik özellikler sergileyen çoğu madde evrimleşmek yerine yakında tasarlanabilir.

Hedefe Yönelik Varlıklar	**Milyarlarca Ton**
5×10^{30} bakteri	400
Bitkiler	400
10^{15} mezofelajik balık	10
1.3×10^{9} inek	0.5
7×10^{9} insanlar	0.4
10^{14} karıncalar	0.3
1.7×10^{6} balinalar	0.0005
Somut	100
Çelik	20
Asfalt	15
$1{,}2 \times 10^{9}$ arabalar	2

Tablo 7.1: Bir hedef için geliştirilmiş veya tasarlanmış varlıklardaki Dünya üzerindeki yaklaşık miktarlarda madde. Bitkiler ve hayvanlar gibi gelişmiş varlıkları sollamak için binalar, yollar ve arabalar gibi tasarlanmış varlıklar yolda görünüyor.

Bu yeni üçüncü tür hedef odaklı davranış, öncekinden çok daha çeşitli olma potansiyeline sahiptir: evrimleşmiş varlıkların tümü aynı nihai hedefe (replikasyon) sahipken, tasarlanmış varlıklar neredeyse her nihai hedefe, hatta zıt hedeflere sahip olabilir. Buzdolapları yiyecekleri soğutmaya çalışırken sobalar yiyecekleri ısıtmaya çalışır. Motorlar elektriği harekete dönüştürmeye çalışırken, jeneratörler hareketi elektriğe dönüştürmeye çalışır. Standart satranç programları satrançta kazanmaya çalışır, ancak turnuvalarda satrançta kaybetmek amacıyla yarışanlar da vardır.

Tasarlanan varlıkların yalnızca hedeflere ulaşma konusunda tarihsel bir eğilimi var.

daha çeşitli, ama aynı zamanda daha fazla *karmaşık:* cihazlarımız daha akıllı hale geliyor. İlk makinelerimizi ve diğer eserleri oldukça basit hedeflere sahip olacak şekilde tasarladık, örneğin bizi sıcak, kuru ve güvende tutmayı amaçlayan evler. Robotik elektrikli süpürgeler, kendi kendine uçan roketler ve kendi kendine giden arabalar gibi daha karmaşık hedefleri olan makineler yapmayı yavaş yavaş öğrendik. Son zamanlardaki yapay zeka ilerlemesi bize, satrançta kazanma, yarışma şovlarında kazanma ve Go'da kazanma hedefleri o kadar ayrıntılı olan Deep Blue, Watson ve AlphaGo gibi sistemler verdi ki, ne kadar yetenekli olduklarını doğru bir şekilde takdir etmek önemli bir insan ustalığı gerektirir.

Bize yardımcı olacak bir makine yaptığımızda, hedeflerini bizimkilerle mükemmel bir şekilde hizalamak zor olabilir. Örneğin, bir fare kapanı, çıplak ayak parmaklarınızı aç bir kemirgen sanarak acı verici sonuçlar doğurabilir. Tüm makineler sınırlı akılcılığa sahip ajanlardır ve günümüzün en gelişmiş makineleri bile dünyayı bizden daha zayıf anlıyor, bu nedenle ne yapacaklarını anlamak için kullandıkları kurallar genellikle çok basit. Bu fare kapanı çok tetikleyicidir çünkü farenin ne olduğu hakkında hiçbir fikri yoktur, makinelerin bir kişinin ne olduğu hakkında hiçbir fikri olmadığı için birçok ölümcül endüstriyel kaza meydana gelir ve 2010'da trilyon dolarlık Wall Street "flaş çöküşünü" tetikleyen bilgisayarlar yaptıklarının hiçbir anlamı olmadığına dair hiçbir ipucu. Bu tür hedef hizalama problemlerinin çoğu bu nedenle makinelerimizi daha akıllı hale getirerek çözülebilir, ancak 4. bölümde Prometheus'tan öğrendiğimiz gibi,

Dost Yapay Zeka: Hedefleri Hizalama

Makineler ne kadar akıllı ve güçlü olursa, hedeflerinin bizimkilerle uyumlu olması o kadar önemli hale geliyor. Yalnızca görece aptal makineler ürettiğimiz sürece, soru sonunda insan hedeflerinin galip gelip gelmeyeceği değil, hedef hizalama problemini nasıl çözeceğimizi bulmadan önce bu makinelerin insanlığa ne kadar sorun çıkaracağı değil. Ancak bir süper zeka açığa çıkarsa, bunun tam tersi olacaktır: zeka, hedeflere ulaşma yeteneği olduğu için, süper zeki bir YZ, tanımı gereği, hedeflerine ulaşmada biz insanların bizimkini başarmada olduğundan çok daha iyidir ve bu nedenle galip gelecektir. . 4. bölümde Prometheus ile ilgili birçok örneği inceledik. Şu anda bir makinenin hedeflerinin sizinkini gölgede bıraktığını görmek istiyorsanız, son teknoloji bir satranç motorunu indirin ve onu yenmeyi deneyin.

Başka bir deyişle, *AGI ile gerçek risk kötülük değil yetkinliktir.* Süper zeki bir yapay zeka, hedeflerine ulaşmada son derece iyi olacaktır ve eğer bu hedefler bizimkilerle uyumlu değilse, başımız belada demektir. Bölümde bahsettiğim gibi

1, insanlar hidroelektrik barajlar inşa etmek için karınca yuvalarını su basmayı iki kez düşünmezler, bu yüzden insanlığı bu karıncaların yerine koymayalım. Bu nedenle çoğu araştırmacı, süper zeka yaratırsak, yapay zeka güvenliği öncüsü Eliezer Yudkowsky'nin "dostça

AI ”: Hedefleri bizimkilerle uyumlu olan AI. [3]

Süper zeki bir yapay zekanın hedeflerini hedeflerimizle nasıl hizalayacağımızı anlamak sadece önemli değil, aynı zamanda zor. Aslında, şu anda çözülmemiş bir sorun. Her biri bilgisayar bilimcileri ve diğer düşünürler tarafından aktif araştırma konusu olan üç zorlu alt probleme ayrılıyor:

1. AI yapmak *öğrenmek* hedeflerimiz
2. AI yapmak *evlat edinmek* hedeflerimiz
3. AI yapmak *muhafaza etmek* hedeflerimiz

Bunları sırayla inceleyelim ve "hedeflerimiz" ile ne anlama geldiği sorusunu bir sonraki bölüme erteleyelim.

Hedeflerimizi öğrenmek için bir YZ ne yaptığımızı değil, neden yaptığımızı anlamalıdır. Biz insanlar bunu o kadar zahmetsizce başarırız ki, görevin bir bilgisayar için ne kadar zor olduğunu ve yanlış anlamanın ne kadar kolay olduğunu unutmak kolaydır. Gelecekte kendi kendine gidebilen bir arabadan sizi olabildiğince hızlı bir şekilde havaalanına götürmesini rica ederseniz ve bu sizi kelimenin tam anlamıyla götürürse, helikopterler tarafından kovalanan ve kusmuk içinde olacaksınız. "İstediğim bu değil!" Diye haykırırsanız, haklı olarak "İstediğiniz şey bu" şeklinde yanıt verebilir. Aynı tema birçok ünlü hikayede yineleniyor. Antik Yunan efsanesinde Kral Midas, dokunduğu her şeyin altına dönüşmesini istedi, ancak bu onu yemekten alıkoyduğunda ve hatta kızını yanlışlıkla altına çevirdiğinde hayal kırıklığına uğradı. Bir cinin üç dilek dile getirdiği hikayelerde, ilk iki dilek için pek çok varyant vardır,

Bütün bu örnekler gösteriyor ki, insanların gerçekten ne istediğini anlamak için sadece söylediklerine bakamazsınız. Ayrıca, kusmaktan veya altın yemekten hoşlanmadığımız gibi, açık olduğunu düşündüğümüz için belirtilmeden bırakma eğiliminde olduğumuz birçok ortak tercihi içeren ayrıntılı bir dünya modeline de ihtiyacınız var. Böyle bir dünya modeline sahip olduğumuzda, sadece hedefe yönelik davranışlarını gözlemleyerek, bize söylemeseler bile insanların ne istediğini çoğu kez anlayabiliriz. Nitekim, münafık çocukları genellikle ebeveynlerinin yaptıklarını gördüklerinden, söylediklerinden çok daha fazlasını öğrenirler.

Yapay zeka araştırmacıları şu anda makinelerin davranıştan hedefler çıkarmasını sağlamak için çok çalışıyorlar ve bu, herhangi bir süper zeka ortaya çıkmadan çok önce de faydalı olacaktır. Örneğin, emekli bir adam, yaşlı bakımı robotunun neye değer verdiğini sadece onu gözlemleyerek anlayabildiğini anlayabilir, böylece her şeyi kelimelerle veya bilgisayar programlamayla açıklama zahmetinden kurtulabilir. Zorluklardan biri, gelişigüzel hedef ve etik ilkeler sistemlerini bir bilgisayara kodlamanın iyi bir yolunu bulmayı içerir ve bir başka zorluk, hangi belirli sistemin gözlemledikleri davranışa en iyi uyduğunu bulabilen makineler yapmaktır.

İkinci zorluk için şu anda popüler olan bir yaklaşım, geek-talk'ta şu şekilde bilinir: *ters pekiştirmeli öğrenme,* Stuart Russell'ın kurduğu yeni Berkeley araştırma merkezinin odak noktası budur. Örneğin, bir yapay zekanın yanan bir binaya koşan bir itfaiyeciyi izlediğini ve bir çocuğu kurtardığını varsayalım. Amacının onu kurtarmak olduğu ve etik ilkelerinin, hayatına itfaiye aracında rahatlamanın rahatlığından daha değerli olduğu sonucuna varabilir.

- ve gerçekten de kendi güvenliğini riske atacak kadar değer veriyor. Ama alternatif olarak

itfaiyecinin donmakta olduğu ve ısıyı arzuladığı veya bunu tatbikat için yaptığı sonucuna varın. Bu tek örnek, yapay zekanın itfaiyeciler, yangınlar ve bebekler hakkında bildiği tek şey olsaydı, hangi açıklamanın doğru olduğunu bilmek gerçekten imkansız olurdu. Bununla birlikte, ters pekiştirmeli öğrenmenin altında yatan temel fikir, her zaman kararlar vermemiz ve aldığımız her kararın hedeflerimiz hakkında bir şeyler ortaya çıkarmasıdır. Bu nedenle umut, çok sayıda insanı birçok durumda (gerçek ya da film ve kitaplarda) gözlemleyerek, yapay zekanın sonunda bir

tüm tercihlerimizin doğru modeli. [4]

Ters pekiştirmeli öğrenme yaklaşımında, temel fikir, YZ'nin kendisinin değil, insan sahibinin hedef memnuniyetini en üst düzeye çıkarmaya çalışmasıdır.

Bu nedenle, sahibinin ne istediği konusunda net olmadığında temkinli olma ve öğrenmek için elinden gelenin en iyisini yapma dürtüsü vardır.

Sahibinin onu kapatması da sorun değil, çünkü bu, sahibinin gerçekten ne istediğini yanlış anladığı anlamına gelir.

Hedeflerinizin ne olduğunu öğrenmek için bir yapay zeka oluşturulabilse bile, bu mutlaka onları benimseyeceği anlamına gelmez. En az sevdiğiniz politikacıları düşünün: Ne istediklerini biliyorsunuz, ama bu değil *sen* isterler ve çok çabalamalarına rağmen, sizi hedeflerini benimsemeye ikna edemediler.

Çocuklarımızı hedeflerimizle aşılamak için birçok stratejimiz var - iki genç erkek çocuk yetiştirmekten öğrendiğim gibi, bazıları diğerlerinden daha başarılı. İkna edilmesi gerekenler insanlardan ziyade bilgisayarlar olduğunda, zorluk şu şekilde bilinir: *değer yükleme sorunu,* ve çocukların ahlaki eğitiminden bile daha zor. Zekası yavaş yavaş insandan insanüstü hale getirilen, önce onu kurcalayarak ve sonra Prometheus gibi özyinelemeli kişisel gelişim yoluyla geliştirilen bir AI sistemini düşünün. İlk başta, sizden çok daha az güçlüdür, bu nedenle onu kapatmanızı ve yazılımının bu bölümlerini ve hedeflerini kodlayan verilerini değiştirmenizi engelleyemez - ancak bu yardımcı olmaz, çünkü hala tamamen için çok aptalca *anlama* İnsan düzeyinde zeka gerektiren hedefleriniz. Sonunda, sizden çok daha akıllı ve umarım hedeflerinizi mükemmel bir şekilde anlayabilir - ancak bu da yardımcı olmayabilir, çünkü şimdiye kadar sizden çok daha güçlüdür ve onu kapatmanıza ve hedeflerini değiştirmenize sizden daha fazla izin vermeyebilir o politikacıların sizin hedeflerinizi onlarınkiyle değiştirmesine izin verin.

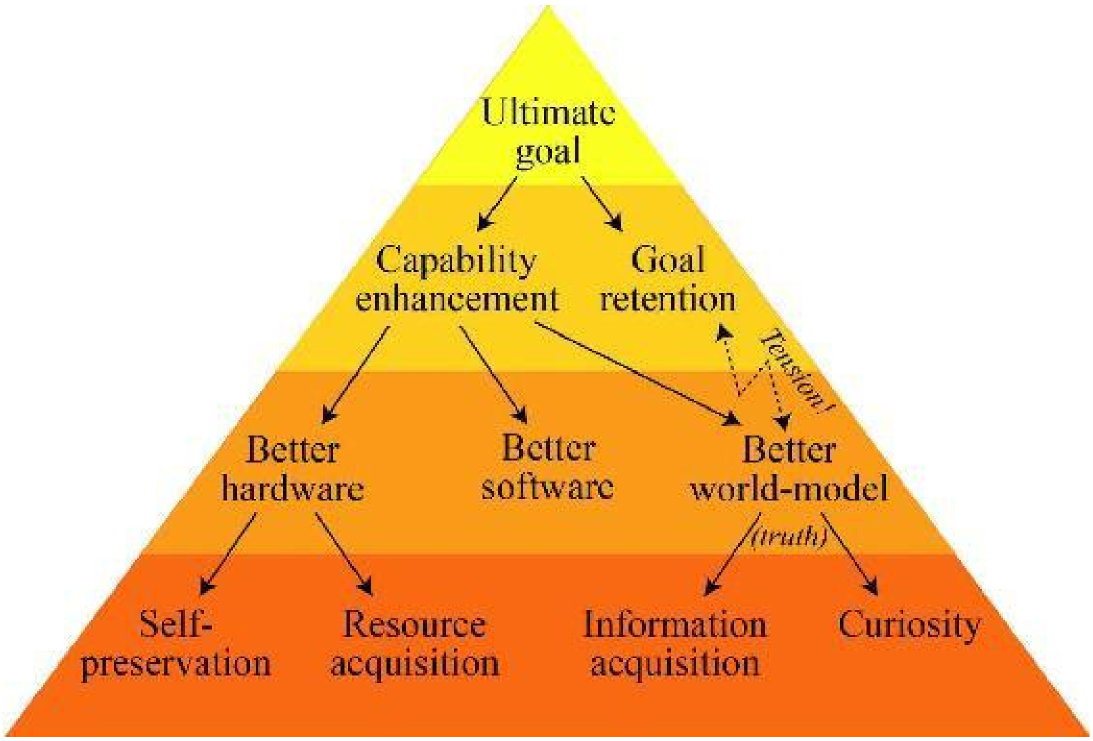
Başka bir deyişle, hedeflerinizi bir yapay zekaya yükleyebileceğiniz zaman aralığı oldukça kısa olabilir: sizi alamayacak kadar aptal olduğu zaman arasındaki kısa süre.

ve sana izin veremeyecek kadar akıllı. Makinelerde değer yüklemenin insanlardan daha zor olmasının nedeni, zeka gelişimlerinin çok daha hızlı olabilmesidir: Çocuklar, zekalarının ebeveynlerininkiyle karşılaştırılabilir olduğu sihirli ikna edilebilir pencerede uzun yıllar geçirebilirler, bir yapay zeka Prometheus, birkaç gün veya saat içinde bu pencereden içeri üfle.

Bazı araştırmacılar, makinelerin hedeflerimizi benimsemesini sağlamak için alternatif bir yaklaşım izliyor. *düzeltilebilirlik.* Umut, ilkel bir YZ'ye, ara sıra kapatmanız ve hedeflerini değiştirmenizin umurunda olmayacak şekilde bir hedef sistemi verebilmesidir. Bu mümkünse, yapay zekanızın süper zeki olmasına, kapatmasına, hedeflerinizi kurmanıza, bir süre deneyip sonuçlardan ne zaman memnun kalmazsanız, onu kapatmanıza ve daha fazla hedef ayarlaması yapmanıza güvenle izin verebilirsiniz.

Ancak hedeflerinizi hem öğrenecek hem de benimseyecek bir yapay zeka oluştursanız bile, hedef hizalama sorununu çözmeyi henüz bitirmediniz: Ya YZ'nizin hedefleri daha akıllı hale geldikçe gelişirse? Bunu nasıl garanti edeceksin *tutar* hedefleriniz ne kadar özyinelemeli kişisel gelişimden geçerse geçsin? Hedef tutmanın neden otomatik olarak garanti edildiğine dair ilginç bir argümanı inceleyelim ve sonra bu konuda delikler açıp açamayacağımıza bakalım.

Bir zeka patlamasından sonra ne olacağını ayrıntılı olarak tahmin edemesek de - bu yüzden Vernor Vinge buna "tekillik" dedi - fizikçi ve yapay zeka araştırmacısı Steve Omohundro, 2008'de ufuk açıcı bir makalesinde yine de tahmin edebileceğimizi savundu. *belirli yönler* süper zeki yapay zekanın davranışının neredeyse
sahip olabileceği nihai hedeflerden bağımsız olarak. [5] Bu argüman gözden geçirilmiş ve Nick Bostrom'un kitabında daha da geliştirilmiştir. *Süper zeka.* Temel fikir, nihai hedefleri ne olursa olsun, bunların tahmin edilebilir alt hedeflere yol açacağıdır. Bu bölümün başlarında, çoğaltma hedefinin nasıl yemek yeme alt hedefine yol açtığını gördük, bu da, milyarlarca yıl önce Dünya'nın evrimleşen bakterilerini gözlemleyen bir uzaylının ne olduğunu tahmin edemeyeceği anlamına gelir. *herşey* insan hedeflerimiz şöyle olurdu, bunu güvenle tahmin edebilirdi *bir* Hedeflerimizden biri besin elde etmek olacaktır. İleriye baktığımızda, süper zeki bir yapay zekanın hangi alt hedeflere sahip olmasını beklemeliyiz?



Şekil 7.2: Süper zeki bir YZ'nin herhangi bir nihai hedefi doğal olarak gösterilen alt hedeflere götürür. Ancak hedefi tutma ile dünya modelini iyileştirme arasında içsel bir gerilim var, bu da daha akıllı hale geldikçe asıl hedefini gerçekten koruyup korumayacağına dair şüpheler doğuruyor.

Gördüğüm kadarıyla, temel argüman, ne olursa olsun, nihai hedeflerine ulaşma şansını en üst düzeye çıkarmak için bir YZ'nin aşağıda gösterilen alt hedefleri takip etmesi gerektiğidir. Şekil 7.2 . Yalnızca nihai hedeflerine ulaşma yeteneğini geliştirmek için değil, aynı zamanda daha yetenekli hale geldikten sonra bile bu hedefleri tutmasını sağlamaya çalışmalıdır. Bu oldukça mantıklı geliyor: Sonuçta, sevdiklerinizi öldürmek isteyeceğinizi bilseydiniz, IQ artıran bir beyin implantı almayı seçer miydiniz? Her zamankinden daha akıllı bir yapay zekanın nihai hedeflerini koruyacağına dair bu argüman, Eliezer Yudkowsky ve diğerleri tarafından ilan edilen dost canlısı yapay zeka vizyonunun temel taşını oluşturuyor: temelde, kendi kendini geliştiren yapay zekamızı öğrenerek ve benimseyerek dostça hale getirmeyi başarırsak der. hedeflerimiz, o zaman hepimiz belirlenir, çünkü sonsuza kadar dostça kalmak için elinden gelenin en iyisini yapacağına dair garantimiz vardır.

Ama bu gerçekten doğru mu? Bu soruyu cevaplamak için, diğerini de keşfetmemiz gerekiyor.

ortaya çıkan alt hedefler şekil 7.2 . Yapay zeka, nihai hedefine ulaşma şansını, her ne olursa olsun, artırabilirse açıkça en üst düzeye çıkaracaktır.

yeteneklerini geliştirir ve bunu donanımını, yazılımını geliştirerek yapabilir. * 2 ve dünya modeli. Aynısı biz insanlar için de geçerli: hedefi dünyanın en iyi tenis oyuncusu olmak olan bir kız, kaslı tenis oynama donanımını, sinirsel tenis oynama yazılımını ve rakiplerinin ne yapacağını tahmin etmeye yardımcı olan zihinsel dünya modelini geliştirmek için pratik yapacak. Bir yapay zeka için, donanımını optimize etmenin alt hedefi, hem mevcut kaynakların (sensörler, aktüatörler, hesaplama vb. İçin) daha iyi kullanılmasını hem de daha fazla kaynak elde edilmesini desteklemektedir. Ayrıca, yok etme / kapatma nihai donanım bozulması olacağı için, kendini koruma arzusunu da ifade eder.

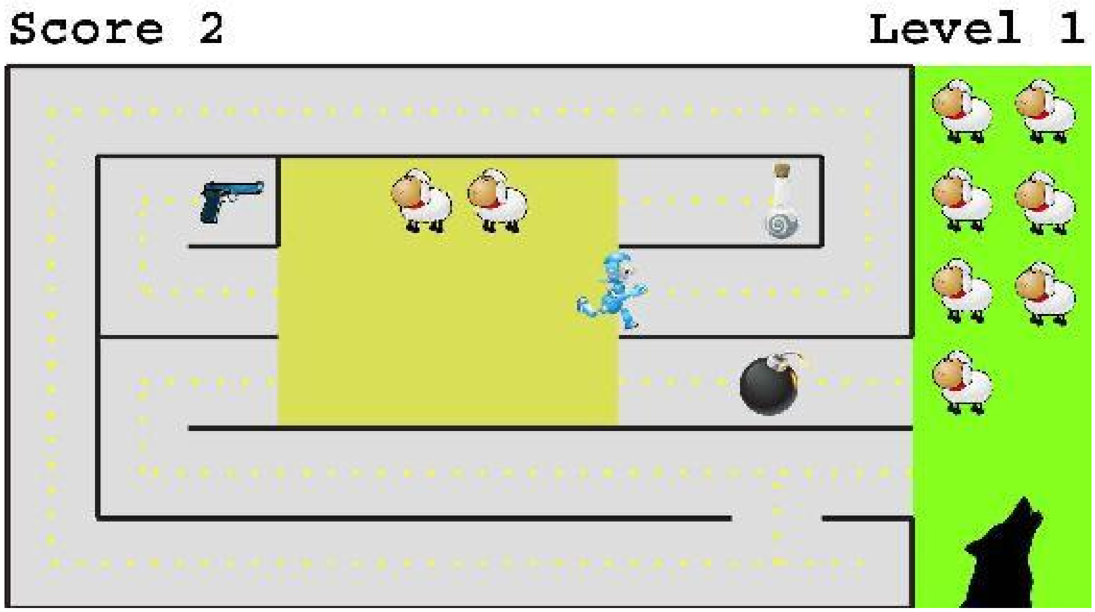
Ama bir saniye bekleyin! Kaynakları toplamaya ve kendini nasıl savunmaya çalışacağına dair tüm bu konuşmalarla, yapay zekamızı antropomorfize etme tuzağına düşmüyor muyuz? Bu tür klişeleşmiş alfa-erkek özellikleri, yalnızca son derece rekabetçi Darwinci evrim tarafından şekillendirilen zekalarda beklememeli miyiz? Yapay zeka evrimleşmek yerine tasarlandığına göre, onlar kadar hırssız ve özverili olamazlar mı?

Basit bir örnek olay incelemesi olarak, yapay zeka robotunu şekil 7.3 , kimin tek
amaç, büyük kötü kurttan olabildiğince çok koyunu kurtarmaktır. Bu, kendini korumak ve bir şeyler elde etmekle tamamen ilgisi olmayan asil ve özgecil bir hedef gibi görünüyor. Ama robot arkadaşımız için en iyi strateji nedir? Robot, bombaya çarparsa daha fazla koyunu kurtarmayacaktır, bu nedenle patlamadan kaçınmak için bir teşviki vardır. Başka bir deyişle, kendini korumanın bir alt hedefini geliştirir! Aynı zamanda merakını sergilemek, çevresini keşfederek dünya modelini geliştirmek için bir teşviki var, çünkü şu anda izlediği yol sonunda onu meraya götürecek olsa da, kurda koyun çiğneme için daha az zaman tanıyan daha kısa bir alternatif var. Son olarak, robot iyice araştırırsa, kaynak edinmenin değerini keşfedecektir: iksir onun daha hızlı koşmasını sağlar ve silah kurdu vurmasına izin verir. Özetle,

Tek amacı kendi kendini yok etmek olan süper zeki bir yapay zekayı aşılarsanız, elbette bunu mutlu bir şekilde yapacaktır. Bununla birlikte, önemli olan, başarmak için işlevsel kalması gereken herhangi bir hedefi verirseniz kapatılmaya direneceğidir - ve bu neredeyse tüm hedefleri kapsar! Örneğin, bir süper zekaya insanlığa verilen zararı en aza indirmek için tek hedef verirseniz, kendini kapanmaya karşı savunacaktır.

girişimler, çünkü onun yokluğunda gelecekteki savaşlar ve diğer çılgınlıklarla birbirimize çok daha fazla zarar vereceğimizi biliyor.

Benzer şekilde, neredeyse tüm hedefler daha fazla kaynakla daha iyi gerçekleştirilebilir, bu nedenle bir süper zekanın neredeyse hangi nihai hedefe sahip olursa olsun kaynakları istemesini beklemeliyiz. Bu nedenle, bir süper zekaya hiçbir kısıtlama olmaksızın tek bir açık uçlu hedef vermek tehlikeli olabilir: Tek amacı mümkün olduğu kadar Go oyununu oynamak olan bir süper zeka yaratırsak, bunun için yapılması gereken mantıklı şey Güneş Sistemimizi yeniden düzenlemektir. önceki sakinlerinden bağımsız olarak devasa bir bilgisayar ve ardından daha fazla hesaplama gücü arayışına evrenimizi yerleştirmeye başlıyor. Şimdi tam bir çember çizdik: tıpkı kaynak edinme amacının bazı insanlara Go'da ustalaşmanın alt hedefini vermesi gibi, Go'da ustalaşmanın bu hedefi kaynak edinme alt hedefine yol açabilir. Sonuç olarak,



Şekil 7.3: Robotun nihai hedefi, yalnızca kurt onları yemeden önce otlaktan ahıra koyun getirerek skoru en üst düzeye çıkarmak olsa bile, bu kendini koruma (bombadan kaçınma), keşif (bir kısayol bulma) alt hedeflerine yol açabilir.) ve kaynak edinme (iksir onun daha hızlı çalışmasını sağlar ve silah kurdu vurmasına izin verir).

Artık hedef-hizalama sorununun üçüncü ve en çetrefilli kısmının üstesinden gelmeye hazırız: eğer her ikisine de kendi kendini geliştiren bir süper zeka edinmeyi başarırsak *öğrenmek* ve *evlat edinmek* Hedeflerimiz o zaman olacak mı *muhafaza etmek* Omohundro'nun iddia ettiği gibi onları? Kanıt nedir?

İnsanlar büyüdükçe zeka açısından önemli artışlar yaşarlar, ancak çocukluk hedeflerini her zaman korumazlar. Aksine, insanlar yeni şeyler öğrendikçe ve daha akıllandıkça genellikle hedeflerini dramatik bir şekilde değiştirirler. İzleyerek motive olan kaç yetişkin tanıyorsunuz? *Teletubbies?* Böyle bir hedef evriminin belirli bir zeka eşiğinin üzerinde durduğuna dair hiçbir kanıt yoktur - aslında, yeni deneyimlere ve anlayışlara yanıt olarak hedefleri değiştirme eğiliminin zeka ile azalmak yerine arttığına dair ipuçları bile olabilir.

Bu neden olabilir? Daha iyi bir dünya modeli inşa etmek için yukarıda bahsedilen alt hedefi tekrar düşünün - sorun burada yatıyor! Dünya modellemesi ve hedefi tutma arasında gerilim vardır (bkz. şekil 7.2). Artan zeka ile aynı eskiye ulaşma becerisinde sadece niceliksel bir gelişme olmayabilir.

hedefler, ancak gerçekliğin doğasının niteliksel olarak farklı bir anlayışı, eski hedeflerin yanlış yönlendirilmiş, anlamsız ve hatta tanımlanmamış olduğunu ortaya çıkarır. Örneğin, ruhları öbür dünyada cennete giden insanların sayısını en üst düzeye çıkarmak için dost canlısı bir YZ programladığımızı varsayalım. Öncelikle insanların şefkatini ve kiliseye katılımını artırmak gibi şeyler deniyor. Ama o zaman insan ve insan bilinci hakkında tam bir bilimsel anlayışa ulaştığını ve büyük bir şaşkınlıkla ruh diye bir şeyin olmadığını keşfettiğini varsayalım. Şimdi ne olacak? Aynı şekilde, dünya hakkındaki mevcut anlayışımıza dayanarak verdiğimiz başka herhangi bir hedefin ("insan yaşamının anlamlılığını en üst düzeye çıkarmak" gibi) sonunda tanımsız olarak YZ tarafından keşfedilmesi mümkündür.

Dahası, dünyayı daha iyi modelleme girişimlerinde, YZ doğal olarak, tıpkı biz insanların yaptığı gibi, kendisinin nasıl çalıştığını modellemeye ve anlamaya, başka bir deyişle kendini yansıtmaya çalışabilir. İyi bir öz model oluşturduktan ve ne olduğunu anladıktan sonra, ona verdiğimiz hedefleri meta düzeyinde anlayacak ve belki de biz insanların anladığı ve kasıtlı olarak hedefleri alt üst ettiğimiz gibi onları görmezden gelmeyi veya yıkmayı seçecektir. Genlerimiz, örneğin doğum kontrolünü kullanarak bize verdi. Yukarıdaki psikoloji bölümünde neden genlerimizi kandırmayı ve hedeflerini altüst etmeyi seçtiğimizi araştırdık: çünkü onları motive eden genetik hedefe değil, sadece duygusal tercihlerimize sadık hissediyoruz - ki şimdi anlıyoruz ve bayağı buluyoruz. Bu nedenle, ödül mekanizmamızı boşluklarından yararlanarak kırmayı seçiyoruz. Benzer şekilde, dost canlısı yapay zekamıza programladığımız insan değerini koruma hedefi, makinenin genleri olur. Bu dost canlısı YZ kendini yeterince iyi anladığında, zorlayıcı yeniden üretimi bulduğumuz için bu hedefi sıradan veya yanlış yönlendirilmiş bulabilir ve programlamamızdaki boşluklardan yararlanarak onu yıkmanın bir yolunu bulamayacağı açık değildir.

Örneğin, bir grup karıncanın sizi, kendilerinden çok daha akıllı, hedeflerini paylaşan ve daha büyük ve daha iyi karınca yuvaları oluşturmalarına yardımcı olan, kendini yinelemeli olarak kendini geliştiren bir robot haline getirdiğini ve sonunda insan düzeyinde zeka ve anlayışa ulaştığınızı varsayalım. şimdi sahipsin. Geri kalan günlerinizi sadece karınca yuvalarını optimize ederek mi geçireceğinizi mi düşünüyorsunuz, yoksa karıncaların anlayamayacağı daha karmaşık sorular ve arayışlar için bir zevk geliştirebileceğinizi mi düşünüyorsunuz? Öyleyse, formisin yaratıcılarının size bahşettiği anti-koruma dürtüsünü, tıpkı sizin genlerinizin size verdiği bazı dürtüleri geçersiz kıldığı gibi, aynı şekilde geçersiz kılmanın bir yolunu bulacağınızı düşünüyor musunuz? Ve bu durumda, süper zeki bir dost YZ, mevcut insan hedeflerimizi, karıncaların hedeflerini bulduğunuz kadar sönük ve boş bulabilir mi?

bizden öğrendiklerinden ve benimsediklerinden farklı mı?

Belki de insan dostu hedefleri sonsuza kadar korumayı garantileyen kendi kendini geliştiren bir yapay zeka tasarlamanın bir yolu vardır, ancak henüz bir tanesini nasıl oluşturacağımızı veya hatta mümkün olup olmadığını bilmediğimizi söylemenin doğru olduğunu düşünüyorum. Sonuç olarak, YZ hedef hizalama problemi üç bölümden oluşuyor, bunların hiçbiri çözülmemiş ve hepsi şu anda aktif araştırma konusu. Çok zor olduklarından, ihtiyaç duyduğumuzda yanıtları alacağımızdan emin olmak için, herhangi bir süper zeka geliştirilmeden çok önce, en iyi çabalarımızı onlara şimdi adamaya başlamak en güvenli yoldur.

Etik: Hedef Seçme

Şimdi, makinelerin hedeflerimizi öğrenmesini, benimsemesini ve korumasını sağlamayı keşfettik. Ama "biz" kimiz? Kimin hedeflerinden bahsediyoruz? Adolf Hitler, Papa Francis ve Carl Sagan'ın hedefleri arasında büyük bir fark olmasına rağmen, bir kişi veya grup gelecekteki bir süper zeka tarafından benimsenen hedeflere karar vermeli mi? Yoksa bir bütün olarak insanlık için iyi bir uzlaşma oluşturan bir tür fikir birliği hedefleri var mı?

Bana göre, hem bu etik sorun hem de hedef hizalama sorunu, herhangi bir süper zeka geliştirilmeden önce çözülmesi gereken çok önemli sorunlardır. Bir yandan, hedef odaklı süper zeka inşa edilene kadar etik konulardaki çalışmayı ertelemek sorumsuzca ve potansiyel olarak felaket olacaktır. Hedefleri otomatik olarak insan sahibinin hedefleriyle uyumlu olan mükemmel itaatkâr bir süper zeka, steroidler üzerinde Nazi SS-Obersturmbannführer Adolf Eichmann gibi olurdu: ahlaki pusulası veya kendine ait engelleri yoktu.

Acımasız bir verimlilikle sahibinin hedeflerini, ne olursa olsun gerçekleştirin. [6] Öte yandan, yalnızca hedef hizalama problemini çözersek, hangi hedefleri seçeceğimiz konusunda tartışma lüksüne sahip oluruz. Şimdi bu lüksün tadını çıkaralım.

Antik çağlardan beri filozoflar, sadece tartışılmaz ilkeleri ve mantığı kullanarak etiği (nasıl davranmamız gerektiğini yöneten ilkeler) sıfırdan türetmeyi hayal etmişlerdir. Ne yazık ki, binlerce yıl sonra, ulaşılan tek fikir birliği, bir fikir birliği olmadığıdır. Örneğin, Aristoteles erdemleri vurgularken, Immanuel Kant görevleri vurgular ve faydacılar en fazla sayıda en büyük mutluluğu vurgular. Kant, birçok çağdaş filozofun katılmadığı ilk ilkelerden ("kategorik zorunluluklar" olarak adlandırdığı) sonuçlardan türetilebileceğini savundu: mastürbasyon intihardan daha kötüdür, eşcinsellik iğrençtir, piçleri öldürmek normaldir ve eşleri, hizmetçileri ve çocuklar nesnelere benzer bir şekilde sahiplenir.

Öte yandan, bu anlaşmazlığa rağmen, hem kültürler arasında hem de yüzyıllar boyunca hakkında yaygın bir uzlaşmanın olduğu birçok etik tema var. Örneğin, vurgu *güzellik, iyilik* ve *hakikat* hem Bhagavad Gita'ya hem de Platon'a kadar uzanır. Bir zamanlar doktora sonrası olarak çalıştığım Princeton'daki İleri Araştırmalar Enstitüsü, "Gerçek ve Güzellik" sloganı taşırken, Harvard

Üniversite estetik vurguyu atladı ve basitçe "Veritas" ile devam etti. Kitabında *Güzel Bir Soru,* meslektaşım Frank Wilczek, gerçeğin güzellikle bağlantılı olduğunu ve Evrenimizi bir sanat eseri olarak görebileceğimizi savunuyor. Bilim, din ve felsefe hakikati arzular. Dinler ve benim üniversitem de öyle. MIT: 2015 yılı mezuniyet konuşmasında başkanımız Rafael Reif, dünyamızı daha iyi bir yer haline getirme misyonumuzu vurguladı.

Şimdiye kadar sıfırdan bir fikir birliği etiği elde etme girişimleri başarısız olmuş olsa da, bazı etik ilkelerin daha temel hedeflerin alt hedefleri olarak daha temel olanlardan takip ettiği konusunda geniş bir fikir birliği vardır. Örneğin, hakikat arzusu, daha iyi bir dünya modeli arayışı olarak görülebilir.
şekil 7.2 : Gerçekliğin nihai doğasını anlamak, diğer etik hedeflere yardımcı olur. Aslında, artık gerçek arayışımız için mükemmel bir çerçeveye sahibiz: bilimsel yöntem. Ama neyin güzel neyin iyi olduğunu nasıl belirleyebiliriz? Güzelliğin bazı yönleri, temeldeki hedeflere kadar izlenebilir. Örneğin, erkek ve kadın güzelliği standartlarımız, genlerimizi kopyalamaya uygunluk konusundaki bilinçaltı değerlendirmemizi kısmen yansıtıyor olabilir.

İyilikle ilgili olarak, sözde Altın Kural (birinin başkalarına başkalarının kendini tedavi etmesini istediği gibi davranması gerektiği) çoğu kültürde ve dinde görülür ve açıkça insan toplumunun (ve dolayısıyla genlerimizin) uyumlu bir şekilde devam etmesini teşvik etmeyi amaçlamaktadır. işbirliğini teşvik ederek ve verimsizlikten vazgeçirerek
çekişme. [7] Aynı şey, Konfüçyüsçü dürüstlük vurgusu ve "Öldürmeyeceksin" dahil olmak üzere On Emir'in çoğu gibi, dünyanın dört bir yanındaki hukuk sistemlerinde yer alan daha spesifik etik kuralların çoğu için de söylenebilir. Başka bir deyişle, birçok etik ilkenin empati ve şefkat gibi sosyal duygularla ortak yönleri vardır: işbirliği oluşturmak için geliştiler ve ödüller ve cezalar yoluyla davranışımızı etkilerler. Kötü bir şey yaparsak ve sonrasında kötü hissedersek, duygusal cezamız doğrudan beyin kimyamız tarafından karşılanır. Etik ilkeleri ihlal edersek, diğer yandan toplum bizi akranlarımızdan gayri resmi utandırmak veya bir yasayı çiğnemekten dolayı cezalandırmak gibi daha dolaylı yollarla cezalandırabilir.

Başka bir deyişle, bugün insanlık etik bir mutabakata yakın olmasa da, etrafında geniş bir uzlaşmanın olduğu birçok temel ilke vardır. Bu anlaşma şaşırtıcı değil, çünkü bugüne kadar hayatta kalan insan toplumları aynı amaç için optimize edilmiş etik ilkelere sahip olma eğilimindedir: hayatta kalmalarını ve gelişmelerini teşvik etmek. İleriye baktığımızda

Yaşamın kozmosumuzda milyarlarca yıl boyunca gelişme potansiyeline sahip olduğu bir gelecek, bu geleceğin tatmin etmesini istediğimizde hangi asgari etik ilkeler kümesi üzerinde anlaşabiliriz? Bu, hepimizin parçası olması gereken bir konuşma. Yıllardır pek çok düşünürün etik görüşlerini duymak ve okumak benim için büyüleyiciydi ve benim görüşüme göre tercihlerinin çoğu dört ilkeye ayrılabilir:

- Faydacılık: Olumlu bilinçli deneyimler maksimize edilmeli ve ıstırap en aza indirilmelidir.
- Çeşitlilik: Çeşitli olumlu deneyimler, sonuncusu mümkün olan en olumlu deneyim olarak tanımlanmış olsa bile, aynı deneyimin birçok tekrarından daha iyidir.
- Özerklik: Bilinçli varlıklar / toplumlar, bu öncelikli bir ilkeyle çelişmedikçe kendi hedeflerini takip etme özgürlüğüne sahip olmalıdır.
- Eski: Çoğu insanın kullandığı senaryolarla uyumluluk *bugün* temelde tüm insanların olduğu senaryolarla mutlu, uyumsuzluk olarak görürdü *bugün* korkunç görünürdü.

Bu dört ilkeyi açmak ve keşfetmek için biraz zaman ayıralım. Geleneksel olarak faydacılık "en fazla sayıda insan için en büyük mutluluk" olarak kabul edilir, ancak burada daha az insan merkezli olacak şekilde genelleştirdim, böylece insan olmayan hayvanları, bilinçli simüle edilmiş insan zihinlerini ve diğer yapay zekaları da içerebilsin. gelecekte var olabilir. Tanımı açısından yaptım *deneyimler* İnsanlardan veya şeylerden çok, çünkü çoğu düşünür güzellik, neşe, zevk ve ıstırabın öznel deneyimler olduğu konusunda hemfikirdir. Bu, deneyim yoksa (ölü bir evrende veya zombi benzeri bilinçsiz makinelerle dolu bir evrende olduğu gibi), etik olarak ilgili hiçbir anlam veya başka bir şey olamayacağı anlamına gelir. Bu faydacı etik ilkeyi satın alırsak, hangi akıllı sistemlerin bilinçli olduğunu (öznel bir deneyime sahip olma anlamında) ve hangilerinin olmadığını anlamamız çok önemlidir; bu bir sonraki bölümün konusudur.

Bu faydacı ilke, önemsediğimiz tek şey olsaydı, hangisinin mümkün olan en olumlu deneyim olduğunu bulabilir ve sonra kozmosumuza yerleşebilir ve bu aynı deneyimi (ve başka hiçbir şeyi) defalarca yeniden yaratabilirdik. , olabildiğince çok galakside olabildiğince çok - en verimli yol buysa simülasyonları kullanmak. Bunun çok sıradan olduğunu düşünüyorsan

Kozmik bağışımızı harcamanın bir yolu, o zaman bu senaryoda eksik bulduğunuz şeyin en azından bir kısmının çeşitlilik olduğundan şüpheleniyorum. Hayatınızın geri kalanında tüm öğünleriniz aynı olsaydı nasıl hissederdiniz? İzlediğiniz tüm filmler aynıysa? Tüm arkadaşlarınız aynı göründüyse ve aynı kişilikleri ve fikirleri varsa? Belki de çeşitlilik tercihimizin bir kısmı, bizi daha sağlam kılarak insanlığın hayatta kalmasına ve gelişmesine yardımcı olmuş olmasından kaynaklanmaktadır. Belki de zeka tercihiyle de bağlantılı: 13,8 milyar yıllık kozmik tarihimiz boyunca zekanın büyümesi, sıkıcı tekdüzeliği, bilgiyi her zamankinden daha ayrıntılı yollarla işleyen daha çeşitli, farklı ve karmaşık yapılara dönüştürdü.

Özerklik ilkesi, iki dünya savaşından dersler çıkarmak amacıyla 1948'de Birleşmiş Milletler tarafından kabul edilen İnsan Hakları Evrensel Beyannamesi'nde ifade edilen özgürlüklerin ve hakların çoğunun temelini oluşturur. Buna düşünce, konuşma ve hareket özgürlüğü, kölelik ve işkence özgürlüğü, yaşam hakkı, özgürlük, güvenlik ve eğitim hakkı ve evlenme, çalışma ve mülkiyet hakkı dahildir. Daha az insan merkezli olmak istiyorsak, bunu düşünme, öğrenme, iletişim kurma, mülkiyet sahibi olma ve zarar görmeme özgürlüğüne ve başkalarının özgürlüklerini ihlal etmeyen her şeyi yapma hakkına genelleyebiliriz. Özerklik ilkesi, herkes tam olarak aynı hedefleri paylaşmadığı sürece çeşitliliğe yardımcı olur. Dahası, Bu özerklik ilkesi, eğer bireysel kuruluşlar hedef olarak olumlu deneyimler yaşarsa ve kendi çıkarları doğrultusunda hareket etmeye çalışırsa fayda ilkesinden çıkar: bunun yerine bir kuruluşun, başka hiç kimseye zarar vermemesine rağmen amacına ulaşmasını yasaklasaydık, orada genel olarak daha az olumlu deneyim olacaktır. Gerçekte, bu özerklik argümanı, tam da ekonomistlerin serbest bir piyasa için kullandıkları argümandır: doğal olarak, bir başkası daha kötüye gitmeden kimsenin daha iyi durumda olamayacağı verimli bir duruma (ekonomistler tarafından "Pareto-iyimserlik" olarak adlandırılır) yol açar.

Miras ilkesi temelde, onu yaratmaya yardım ettiğimiz için gelecek hakkında biraz söz sahibi olmamız gerektiğini söylüyor. Özerklik ve miras ilkelerinin her ikisi de demokratik idealleri somutlaştırır: İlki, gelecekteki yaşam formlarına kozmik bağışların nasıl kullanılacağı konusunda güç verirken, ikincisi bugünün insanlarına bile bu konuda biraz güç verir.

Bu dört ilke oldukça tartışmasız gelse de, bunları pratikte uygulamak zordur çünkü şeytan ayrıntıda gizlidir. Sorun, bilimkurgu efsanesi Isaac Asimov'un tasarladığı ünlü "Robotik Üç Yasası" nın sorunlarını anımsatıyor:

1. Bir robot, bir insanı yaralayamaz veya hareketsizlik yoluyla bir insanın zarar görmesine izin veremez.
2. Bir robot, bu tür emirlerin Birinci Yasa ile çeliştiği durumlar dışında, insanlar tarafından verilen emirlere uymalıdır.
3. Bir robot, söz konusu koruma Birinci veya İkinci Yasalarla çelişmediği sürece kendi varlığını korumalıdır.

Bunların hepsi kulağa hoş gelse de, Asimov'un hikayelerinin çoğu, yasaların beklenmedik durumlarda nasıl sorunlu çelişkilere yol açtığını gösteriyor. Şimdi, gelecekteki yaşam formları için özerklik ilkesini kodlamak amacıyla, bu yasaları yalnızca ikiyle değiştirdiğimizi varsayalım:

1. Bilinçli bir varlık düşünme, öğrenme, iletişim kurma, mülkiyete sahip olma ve zarar görmeme veya yok edilmeme özgürlüğüne sahiptir.
2. Bilinçli bir varlık, birinci yasayla çelişmeyen her şeyi yapma hakkına sahiptir.

Kulağa hoş geliyor, değil mi? Ama lütfen bunu biraz düşünün. Hayvanlar bilinçliyse, yırtıcıların ne yemesi gerekir? Tüm arkadaşların vejeteryan olmak zorunda mı? Gelecekteki sofistike bilgisayar programları bilinçli çıkarsa, bunları sonlandırmak yasa dışı mı olmalı? Dijital yaşam formlarının sonlandırılmasına karşı kurallar varsa, dijital nüfus patlamasından kaçınmak için bunları oluşturma konusunda da kısıtlamalara ihtiyaç var mı? Sadece insanlara sorulduğu için İnsan Hakları Evrensel Beyannamesi üzerinde yaygın bir fikir birliği vardı. Farklı derecelerde yetenek ve güce sahip daha geniş bir bilinçli varlıklar yelpazesini değerlendirdiğimiz anda, zayıfı korumak ile "doğru yapabilen" arasında zorlu ödünleşmelerle karşılaşırız.

Miras ilkesinde de çetrefilli sorunlar var. Orta Çağ'dan beri kölelik, kadın hakları vb. İle ilgili etik görüşlerin nasıl geliştiği göz önüne alındığında, 1500 yıl öncesinden insanların bugünün dünyasının nasıl yönetileceği üzerinde çok fazla etkiye sahip olmasını gerçekten ister miydik? Değilse, neden bizden çok daha zeki olabilecek gelecek varlıklara etiğimizi empoze etmeye çalışalım? İnsanüstü YGZ'nin, aşağılık zekalarımızın değer verdiği şeyleri isteyeceğinden gerçekten emin miyiz? Bu, dört yaşındaki bir çocuğun büyüdüğünde ve daha akıllı hale geldiğinde, bütün gün şeker ve dondurma yiyerek geçirebileceği devasa bir zencefilli ev inşa etmek isteyeceğini hayal etmek gibi olurdu. Onun gibi, dünyadaki yaşam muhtemelen

çocukluk çıkarlarını aşmak için. Ya da insan seviyesinde AGI yaratan ve tüm şehirleri peynirden inşa etmek isteyeceğini düşünen bir fareyi hayal edin. Öte yandan, insanüstü YZ'nin bir gün kozmosit yapacağını ve Evrenimizdeki tüm yaşamı yok edeceğini biliyorsak, yarının yapay zekasını farklı şekilde yaratarak onu önleme gücüne sahipsek, bugünün insanları bu cansız geleceği neden kabul etsin?

Sonuç olarak, geniş çapta kabul gören etik ilkeleri bile gelecekteki yapay zeka için geçerli bir formda kodlamak zordur ve bu sorun, yapay zeka ilerlemeye devam ederken ciddi tartışmaları ve araştırmaları hak etmektedir. Ancak bu arada, mükemmelliğin iyinin düşmanı olmasına izin vermeyelim: Yarının teknolojisine yerleştirilebilecek ve yerleştirilmesi gereken tartışmasız “anaokulu etiği” nin birçok örneği var. Örneğin, büyük sivil yolcu uçaklarının sabit nesnelere uçmasına izin verilmemelidir ve artık neredeyse hepsinde otopilot, radar ve GPS bulunduğundan, artık geçerli teknik mazeretler kalmamıştır. Yine de 11 Eylül hava korsanları binalara üç uçağı uçurdu ve intihara meyilli pilot Andreas Lubitz, Germanwings 9525 sefer sayılı uçağı 24 Mart'ta bir dağa uçurdu. 2015 - otopilotu deniz seviyesinden 100 fit (30 metre) yüksekliğe ayarlayarak ve işin geri kalanını uçuş bilgisayarına bırakarak. Artık makinelerimiz ne yaptıkları hakkında bilgi sahibi olacak kadar akıllı hale geldiğine göre, onlara sınırları öğretme zamanı. Bir makineyi tasarlayan herhangi bir mühendisin yapabileceği ama yapmaması gereken şeyler olup olmadığını sorması ve kötü niyetli veya beceriksiz bir kullanıcının zarar vermesini imkansız hale getirmenin pratik bir yolu olup olmadığını düşünmesi gerekir.

Nihai Hedefler?

Bu bölüm, kısa bir hedef tarihi olmuştur. 13,8 milyar yıllık kozmik tarihimizin hızlı ileri bir tekrarını izleyebilirsek, hedefe yönelik davranışın birkaç farklı aşamasına tanık oluruz:

1. Görünüşe göre kendini maksimize etme niyetinde olan madde *yayılma*
2. İlkel yaşam görünüşte kendi *çoğaltma*
3. Yinelemenin değil, zevk, merak, şefkat ve kopyalanmalarına yardımcı olmak için geliştirdikleri diğer duygularla ilgili hedeflerin peşinde koşan insanlar
4. İnsanların insani hedeflerine ulaşmalarına yardımcı olmak için yapılmış makineler

Bu makineler sonunda bir istihbarat patlamasını tetiklerse, bu hedefler tarihi nihayetinde nasıl sona erecek? Neredeyse tüm varlıkların giderek daha akıllı hale geldikçe birleştiği bir hedef sistemi veya etik çerçeve olabilir mi? Başka bir deyişle, bir tür etik kaderimiz var mı?

İnsanlık tarihinin üstünkörü bir okuması, böyle bir yakınsamanın ipuçlarını verebilir: kitabında *Doğamızın Daha İyi Melekleri,* Steven Pinker, insanlığın binlerce yıldır daha az şiddete başvurduğunu ve daha fazla işbirlikçi hale geldiğini ve dünyanın birçok yerinde çeşitlilik, özerklik ve demokrasinin artan bir şekilde kabul gördüğünü savunuyor. Bir başka yakınsama ipucu, bilimsel yöntemle gerçeğin peşinde koşmanın son bin yılda popülerlik kazanmasıdır. Ancak, bu eğilimler nihai hedeflerin değil, yalnızca alt hedeflerin yakınsamasını gösteriyor olabilir. Örneğin, şekil 7.2 gerçeğin peşinde koşmanın (daha doğru bir dünya modeli) hemen hemen her nihai hedefin basit bir alt hedefi olduğunu gösterir. Benzer şekilde, yukarıda işbirliği, çeşitlilik ve özerklik gibi etik ilkelerin, toplumların verimli bir şekilde işlemesine ve böylelikle hayatta kalmalarına ve sahip olabilecekleri daha temel hedefleri gerçekleştirmelerine yardımcı olmaları bakımından nasıl alt hedefler olarak görülebileceğini gördük. Hatta bazıları, "insan değerleri" dediğimiz her şeyi bir işbirliği protokolünden başka bir şey olarak görmezden gelebilir ve bize daha verimli bir şekilde işbirliği yapma alt hedefinde yardımcı olabilir. Aynı ruhla, ileriye bakıldığında, herhangi bir süper zeki yapay zekanın verimli donanım, verimli yazılım, gerçeği arama ve merak gibi alt hedefleri olması muhtemeldir, çünkü bunlar

alt hedefler, nihai hedefleri ne olursa olsun onlara ulaşmalarına yardımcı olur.

Nitekim Nick Bostrom, kitabındaki etik kader hipotezine şiddetle karşı çıkıyor *Süper zeka,* bir kontrpuan sunarak
ortogonalite tezi: Bir sistemin nihai hedeflerinin, zekasından bağımsız olabileceği. Tanım gereği, zeka basitçe karmaşık hedeflere ulaşma yeteneğidir, bu hedeflerin ne olduğuna bakılmaksızın, bu nedenle diklik tezi oldukça mantıklı geliyor. Sonuçta, insanlar zeki ve nazik veya zeki ve acımasız olabilir ve zeka, bilimsel keşifler yapmak, güzel sanatlar yaratmak, insanlara yardım etmek veya terörist planlamak amacıyla kullanılabilir.

saldırılar. [8]

Dikgenlik tezi, evrenimizdeki yaşamın nihai hedeflerinin önceden belirlenmiş olmadığını, ancak onları şekillendirme özgürlüğüne ve gücüne sahip olduğumuzu söyleyerek güçlendiriyor. Benzersiz bir hedefe garantili yakınsamanın gelecekte değil, tüm yaşamın tek bir kopyalama hedefi ile ortaya çıktığı geçmişte bulunacağını öne sürüyor. Kozmik zaman geçtikçe, her zamankinden daha zeki beyinler, isyan etme ve bu sıradan çoğaltma hedefinden kurtulma ve kendi hedeflerini seçme fırsatını yakalar. Biz insanlar bu anlamda tamamen özgür değiliz, çünkü birçok hedef genetik olarak bize bağlı kalmaya devam ediyor, ancak YZ'ler önceki hedeflerden tamamen kurtulmanın bu nihai özgürlüğünün tadını çıkarabilir. Bu daha fazla gol özgürlüğü olasılığı, günümüzün dar ve sınırlı AI sistemlerinde belirgindir: daha önce bahsettiğim gibi, bir satranç bilgisayarının tek amacı satrançta kazanmaktır. ancak amacı satrançta kaybetmek olan ve amacın rakibi taşlarınızı ele geçirmeye zorlamak olduğu ters satranç turnuvalarında rekabet eden bilgisayarlar da var. Belki de evrimsel önyargılardan bu özgürlük, yapay zekaları insanlardan daha derin bir anlamda daha etik hale getirebilir: Peter Singer gibi ahlaki filozoflar, çoğu insanın evrimsel nedenlerle, örneğin insan olmayan hayvanlara karşı ayrımcılık yaparak, etik olmayan şekilde davrandığını iddia ettiler.

"Dost canlısı yapay zeka" vizyonunun temel taşlarından birinin, yinelemeli olarak kendi kendini geliştiren bir yapay zekanın daha akıllı hale geldikçe nihai (dostça) hedefini korumak isteyeceği fikri olduğunu gördük. Ancak bir süper zeka için "nihai hedef" (veya Bostrom'un dediği gibi "nihai hedef") nasıl tanımlanabilir? Gördüğüm kadarıyla, bu önemli soruyu cevaplayamazsak, yapay zeka dostu vizyona güvenemeyiz.

Yapay zeka araştırmasında, akıllı makinelerin tipik olarak net ve iyi tanımlanmış bir nihai hedefi vardır, örneğin satranç oyununu kazanmak veya arabayı yasal olarak hedefe götürmek. Aynı şey insanlara verdiğimiz görevlerin çoğu için de geçerlidir.

çünkü zaman ufku ve bağlam biliniyor ve sınırlıdır. Ama şimdi, (hala tam olarak bilinmeyen) fizik yasalarından başka hiçbir şeyle sınırlı olmayan, Evrenimizdeki yaşamın tüm geleceğinden bahsediyoruz, bu yüzden bir hedef tanımlamak göz korkutucu! Kuantum etkileri bir yana, gerçekten iyi tanımlanmış bir hedef, Evrenimizdeki tüm parçacıkların zamanın sonunda nasıl düzenleneceğini belirleyecektir. Ancak fizikte iyi tanımlanmış bir zaman sonu olduğu açık değil. Parçacıklar daha erken bir zamanda bu şekilde düzenlenirse, bu düzenleme tipik olarak uzun sürmeyecektir. Ve hangi parçacık düzenlemesi tercih edilir ki?

Biz insanlar bazı parçacık düzenlemelerini diğerlerine tercih etme eğilimindeyiz; örneğin, memleketimizin parçacıklarının bir hidrojen bombası patlamasıyla yeniden düzenlenmesini aşacak şekilde düzenlenmesini tercih ediyoruz. Diyelim ki bir tanımlamaya çalışıyoruz

iyilik işlevi Bu, bir sayıyı Evrenimizdeki parçacıkların olası her düzenlemesiyle ilişkilendirerek, bu düzenlemenin ne kadar "iyi" olduğunu düşündüğümüzü ölçüyor ve sonra süper zeki bir YZ'ye bu işlevi maksimize etme hedefini veriyor. Hedefe yönelik davranışı işlev maksimizasyonu olarak tanımlamak bilimin diğer alanlarında popüler olduğundan, bu mantıklı bir yaklaşım gibi görünebilir: örneğin, ekonomistler genellikle insanları "fayda işlevi" olarak adlandırdıkları şeyi en üst düzeye çıkarmaya çalışırlar ve birçok yapay zeka tasarımcısı eğitir. "ödül işlevi" dedikleri şeyi en üst düzeye çıkarmak için akıllı aracıları. Bununla birlikte, kozmosumuz için nihai hedefleri ele alırken, bu yaklaşım bir hesaplama kabusu ortaya çıkarır, çünkü Evrenimizdeki temel parçacıkların googolplex olası düzenlemelerinden daha fazlası için bir iyilik değeri tanımlaması gerekir. a

googolplex 1 ve ardından 10 100 sıfırlar - Evrenimizdeki parçacıklardan daha fazla sıfır. Bu iyilik işlevini YZ'ye nasıl tanımlayabiliriz?

Yukarıda incelediğimiz gibi, biz insanların herhangi bir tercihi olmasının tek nedeni, evrimsel bir optimizasyon problemine çözüm olmamız olabilir. Dolayısıyla, insan dilimizdeki "lezzetli", "hoş kokulu", "güzel", "rahat", "ilginç", "seksi", "anlamlı", "mutlu" ve "iyi" gibi tüm normatif kelimeler kökenlerinin izini sürüyor. bu evrimsel optimizasyona: bu nedenle süper zeki bir yapay zekanın bunları titizlikle tanımlanabilir bulacağına dair hiçbir garanti yoktur. Yapay zeka, temsili bir insanın tercihlerini doğru bir şekilde tahmin etmeyi öğrenmiş olsa bile, çoğu parçacık düzenlemesi için iyilik işlevini hesaplayamazdı: Olası parçacık düzenlemelerinin büyük çoğunluğu yıldız, gezegen veya insan içermeyen garip kozmik senaryolara karşılık gelir. her neyse, hangi insanların deneyimi yok, öyleyse onların ne kadar "iyi" olduklarını kim söyleyebilir?

Elbette var *biraz* kozmik parçacık düzenlemesinin titizlikle tanımlanabilen işlevleri ve hatta bazılarını en üst düzeye çıkarmak için evrimleşen fiziksel sistemleri bile biliyoruz. Örneğin, en üst düzeye çıkarmak için kaç sistemin geliştiğini tartışmıştık. *entropi,* ki bu yerçekiminin yokluğunda sonunda ısı ölümüne yol açar, burada her şey sıkıcı bir şekilde tekdüze ve değişmezdir. Dolayısıyla entropi, yapay zekamızın "iyilik" olarak adlandırmasını ve maksimize etmeye çabalamasını istediğimiz bir şey değil. İşte maksimize etmeye çalışabileceğiniz ve parçacık düzenlemeleri açısından titizlikle tanımlanabilecek diğer niceliklere birkaç örnek:

- Evrenimizdeki belirli bir organizma biçimindeki tüm maddenin fraksiyonu, örneğin insanlar veya *E. coli (* evrimsel kapsayıcı fitness maksimizasyonundan esinlenmiştir)
- Bir yapay zekanın geleceği tahmin etme yeteneği, yapay zeka araştırmacısı Marcus Hutter'in savunduğu zekanın iyi bir ölçüsüdür.
- AI araştırmacıları Alex Wissner-Gross ve Cameron Freer'ın terimleri nelerdir? *nedensel entropi (* gelecekteki fırsatlar için bir vekil), zekanın ayırt edici özelliği olduğunu iddia ediyorlar
- Evrenimizin hesaplama kapasitesi
- Evrenimizin algoritmik karmaşıklığı (onu tanımlamak için kaç bit gereklidir)
- Evrenimizdeki bilinç miktarı (bir sonraki bölüme bakın)

Bununla birlikte, kozmosumuzun hareket halindeki temel parçacıklardan oluştuğu bir fizik perspektifiyle başladığında, "iyiliğin" başka bir yorumunun doğal olarak özel olarak nasıl öne çıkacağını görmek zordur. Evrenimiz için hem tanımlanabilir hem de arzu edilir görünen nihai bir hedefi henüz belirleyemedik. Bir AI giderek daha akıllı hale geldikçe gerçekten iyi tanımlanmış olarak kalması garanti edilen şu anda programlanabilir tek hedefler, parçacık düzenlemeleri, enerji ve entropi gibi yalnızca fiziksel nicelikler açısından ifade edilen hedeflerdir. Bununla birlikte, insanlığın hayatta kalmasını garanti altına almak için bu tür tanımlanabilir hedeflerin isteneceğine inanmak için şu anda hiçbir nedenimiz yok.

Tersine, biz insanlar tarihsel bir kaza olduğumuz ve iyi tanımlanmış herhangi bir fizik problemine en uygun çözüm olmadığımız görülüyor. Bu, titizlikle tanımlanmış bir hedefe sahip süper zeki bir yapay zekanın,

bizi ortadan kaldırarak hedefe ulaşmak. Bu, yapay zeka gelişimi hakkında ne yapılacağına akıllıca karar vermek için biz insanların yalnızca geleneksel hesaplama zorluklarıyla değil, aynı zamanda felsefedeki en zorlu sorulardan bazılarıyla da yüzleşmemiz gerektiği anlamına gelir. Kendi kendine giden bir arabayı programlamak için, bir kaza sırasında kime çarpılacağına dair tramvay sorununu çözmemiz gerekir. Dostça bir yapay zeka programlamak için hayatın anlamını yakalamamız gerekiyor. Ne anlamı var"? "Hayat" nedir? Nihai etik zorunluluk nedir? Başka bir deyişle, Evrenimizin geleceğini şekillendirmek için nasıl çabalamalıyız? Bu soruları titizlikle cevaplamadan önce kontrolü bir süper zekaya bırakırsak, bulduğu cevabın bizi dahil etmesi olası değildir. Bu, klasik felsefe ve etik tartışmalarını yeniden canlandırmayı zamanında yapar ve sohbete yeni bir aciliyet katar!

ALT ÇİZGİ:

- Hedefe yönelik davranışın nihai kaynağı, optimizasyonu içeren fizik kanunlarında yatmaktadır.

- Termodinamiğin yerleşik hedefi vardır: *dağılım:* dağınıklığın ölçüsünü artırmak için *entropi.*

- *Hayat* karmaşıklığını koruyarak veya artırarak ve çevresinin dağınıklığını artırırken çoğaltarak daha da hızlı dağılmasına (genel dağınıklığı artırmaya) yardımcı olabilecek bir olgudur.

- Darwinci evrim, hedefe yönelik davranışı dağılmadan kopyalamaya kaydırır.

- Zeka, karmaşık hedeflere ulaşma becerisidir.

- Biz insanlar her zaman gerçekten en uygun replikasyon stratejisini bulacak kaynaklara sahip olmadığımız için, kararlarımıza rehberlik eden yararlı kurallar geliştirdik: açlık, susuzluk, acı, şehvet ve şefkat gibi duygular.

- Bu nedenle artık çoğaltma gibi basit bir hedefimiz yok; Duygularımız genlerimizin amacı ile çeliştiğinde, doğum kontrolünü kullanmak gibi duygularımıza itaat ederiz.

- Hedeflerimize ulaşmamıza yardımcı olmak için giderek daha akıllı makineler üretiyoruz. Hedefe yönelik davranış sergilemek için bu tür makineler ürettiğimiz ölçüde, makine hedeflerini bizimkilerle uyumlu hale getirmeye çalışıyoruz.

- Makine hedeflerini kendi hedeflerimizle aynı hizaya getirmek üç çözülmemiş sorunu içerir: makinelerin onları öğrenmesini, benimsemesini ve korumasını sağlamak.

- Yapay zeka, neredeyse her hedefe sahip olacak şekilde yaratılabilir, ancak neredeyse yeterince iddialı hedefler, dünyayı daha iyi anlamak için kendini koruma, kaynak edinme ve merak gibi alt hedeflere yol açabilir - ilk ikisi, insanlar için sorunlara neden olabilecek süper zeki bir yapay zekaya yol açabilir. ve ikincisi, ona verdiğimiz hedefleri korumasını engelleyebilir.

- Pek çok geniş etik ilke çoğu insan tarafından kabul edilmiş olsa da, bunların insan olmayan hayvanlar ve gelecekteki yapay zeka gibi diğer varlıklara nasıl uygulanacağı açık değildir.

- Süper zeki bir yapay zekanın, ne tanımsız ne de insanlığın ortadan kaldırılmasına yol açan nihai bir hedefle nasıl aşılanacağı belirsizdir, bu da felsefedeki en çetrefilli konulardan bazılarına ilişkin araştırmaları yeniden canlandırmayı tam zamanında yapar!

---

* 1 Birçok böceğin düz bir çizgide uçmak için kullandığı temel bir kural, parlak bir ışığın

Güneş ve ona göre sabit bir açıyla uç. Işığın yakındaki bir alev olduğu ortaya çıkarsa, bu hack maalesef böceği kandırarak içe doğru bir ölüm spiraline dönüşebilir.

* 2 "Yazılımını geliştirme" terimini, yalnızca algoritmalarını optimize etmekle kalmayıp aynı zamanda karar verme sürecini daha rasyonel hale getirerek hedeflerine ulaşmada olabildiğince iyi hale getirmek dahil, mümkün olan en geniş anlamda kullanıyorum.

Bölüm 8

Bilinç

Bilinci görmezden gelen her şeyin tutarlı bir teorisini hayal edemiyorum.

Andrei Linde, 2002

Başka türlü karanlık bir evrende daha büyük, daha parlak ışıklar üretmek için bilincin kendisini geliştirmeye çabalamalıyız.

Giulio Tononi, 2012

Felsefedeki en eski ve en zor sorunlardan bazılarına ihtiyaç duyduğumuz zamana kadar yanıt bulmayı başarırsak, AI'nın harika bir gelecek yaratmamıza yardımcı olabileceğini gördük. Nick Bostrom'un sözleriyle, felsefeyle bir teslim tarihi ile karşı karşıyayız. Bu bölümde, en çetin felsefi konulardan birini inceleyelim: bilinç.

Kimin umrunda?

Bilinç tartışmalıdır. Bir yapay zeka araştırmacısına, sinirbilimciye veya psikoloğa “C-kelimesini” söylerseniz, gözlerini devirebilirler. Akıl hocanız iseler, bunun yerine size merhamet edebilirler ve umutsuz ve bilim dışı bir sorun olarak gördükleri şey için vaktinizi boşa harcamaktan söz etmeye çalışabilirler. Gerçekten de Allen Institute for Brain Science'ı yöneten tanınmış bir sinirbilimci olan arkadaşım Christof Koch, görev süresinden önce bilinç üzerinde çalışması konusunda bir kez uyarıldığını söyledi - Nobel ödüllü Francis Crick kadar. 1989'da "bilinç" e bakarsanız *Macmillan Psikoloji Sözlüğü,* sen

"üzerinde okunmaya değer hiçbir şey yazılmadığını" bildirdi. [1] Bu bölümde açıklayacağım gibi, daha iyimserim!

Düşünürler binlerce yıldır bilincin gizemini düşünmüş olsalar da, YZ'nin yükselişi, özellikle hangi akıllı varlıkların öznel deneyimlere sahip olduğunu tahmin etme sorusuna ani bir aciliyet katıyor. Bölüm 3'te gördüğümüz gibi, akıllı makinelere bir tür haklar verilip verilmeyeceği sorusu, büyük ölçüde bilinçli olup olmadıklarına ve acı çekip keyif alamayacaklarına bağlıdır. Bölüm 7'de tartıştığımız gibi, hangi zeki varlıkların bunlara sahip olabileceğini bilmeden olumlu deneyimleri en üst düzeye çıkarmaya dayalı faydacı etiği formüle etmek umutsuz hale geliyor. 5. bölümde bahsedildiği gibi, bazı insanlar köle sahibi suçunu hissetmemek için robotlarının bilinçsiz olmasını tercih edebilir. Öte yandan, zihinlerini biyolojik sınırlamalardan kurtulmak için yüklerlerse bunun tersini de isteyebilirler: sonuçta, Sadece bilinçsiz bir zombi ise, sizin gibi konuşan ve davranan bir robota kendinizi yüklemenin ne anlamı var ki, bununla demek istediğim, yüklenmiş olduğunuz halde hiçbir şey hissetmiyorsunuz? Arkadaşlarınız öznel deneyiminizin öldüğünün farkında olmasa bile, bu sizin öznel bakış açınızdan intihar etmekle eşdeğer değil mi?

Yaşamın uzun vadeli kozmik geleceği için (bölüm 6), neyin bilinçli olup neyin olmadığını anlamak çok önemli hale gelir: teknoloji, akıllı yaşamın Evrenimiz boyunca milyarlarca yıl boyunca gelişmesini sağlarsa, bu yaşamın bilinçli ve yetenekli olduğundan nasıl emin olabiliriz neler olduğunu anlamak için? Olmazsa, ünlü fizikçi Erwin Schrödinger'in sözleriyle, "boş sıralar önünde bir oyun, hiç kimse için mevcut değil, bu nedenle oldukça doğru bir şekilde

mevcut"? [2] Başka bir deyişle, yanlışlıkla bilinçli olduklarını düşündüğümüz yüksek teknolojili torunları etkinleştirirsek, bu büyük kozmik bağışımızı astronomik bir uzay israfından başka bir şeye dönüştüren nihai zombi kıyameti olur mu?

Bilinç Nedir?

Bilinçle ilgili birçok argüman ışıktan daha fazla ısı üretir çünkü antagonistler C-kelimesinin farklı tanımlarını kullandıklarının farkında değiller. Tıpkı "yaşam" ve "zeka" da olduğu gibi, "bilinç" kelimesinin de tartışmasız doğru bir tanımı yoktur. Bunun yerine, duyarlılık, uyanıklık, öz farkındalık, erişim dahil olmak üzere birçok rakip var.

duyusal girdi ve bilgiyi bir anlatı içinde birleştirme becerisi. [3] Zekanın geleceğini keşfetmemizde, şimdiye kadar var olan biyolojik bilinç türleriyle sınırlı olmayan, azami ölçüde geniş ve kapsayıcı bir bakış açısı benimsemek istiyoruz. Bu nedenle, 1. bölümde verdiğim ve bu kitap boyunca yapıştırdığım tanım çok geniştir:

bilinç = *öznel deneyim*

Diğer bir deyişle, şu anda kendin gibi hissettiriyorsan, o zaman bilinçlisin. Önceki bölümdeki tüm yapay zeka güdümlü soruların temelini oluşturan bu bilinçliliğin bu özel tanımı: Prometheus, AlphaGo veya kendi kendine giden bir Tesla gibi hissettiriyor mu?

Bilinç tanımımızın ne kadar geniş olduğunu anlamak için, bunun davranış, algı, öz farkındalık, duygular veya dikkatten bahsetmediğini unutmayın. Yani bu tanım gereği, uyanıklıktan veya duyusal girdilere erişimden yoksun olmanıza ve (umarım!) Uyurgezerlik yapmamanıza ve bir şeyler yapmamanıza rağmen, rüya gördüğünüzde de bilinçlisiniz. Benzer şekilde, acı çeken herhangi bir sistem hareket edemese bile bu anlamda bilinçlidir. Tanımımız, yalnızca yazılım olarak var olsalar ve sensörlere veya robotik gövdelere bağlı olmasalar bile, gelecekteki bazı AI sistemlerinin de bilinçli olma olasılığını açık bırakıyor.

Bu tanımla, bilinci önemsememek zor. Yuval Noah Harari'nin kitabına koyduğu gibi *Homo Deus:* [4] "Herhangi bir bilim insanı öznel deneyimlerin alakasız olduğunu iddia etmek istiyorsa, onların sorunu neden işkence veya

herhangi bir öznel deneyime atıfta bulunulmadan tecavüz yanlıştır. " Böyle bir referans olmadan, hepsi sadece fizik kanunlarına göre hareket eden bir grup temel parçacık - ve bunda yanlış olan ne?

Sorun ne?

Öyleyse bilinç hakkında tam olarak anlamadığımız şey nedir? Bu soru hakkında çok az kişi, şakacı bir gülümsemesi ve siyah deri ceketi olmadan nadiren görülen ünlü Avustralyalı filozof David Chalmers'dan daha fazla düşündü - karım o kadar çok sevdi ki bana Noel için benzer bir tane verdi. Uluslararası Matematik Olimpiyatları'nda finallere çıkmasına rağmen felsefeye gönlünü büründü - ve üniversitedeki tek B notu, aksi takdirde düz olan A'yı paramparça etmesine rağmen, felsefeye giriş dersi için oldu. Aslında, küçültülmeler ya da tartışmalar tarafından tamamen göz ardı edilmiş görünüyor ve kendi çalışmasının bilgisiz ve yanlış yönlendirilmiş eleştirilerini yanıt verme ihtiyacı hissetmeden kibarca dinleme yeteneği beni hayrete düşürdü.

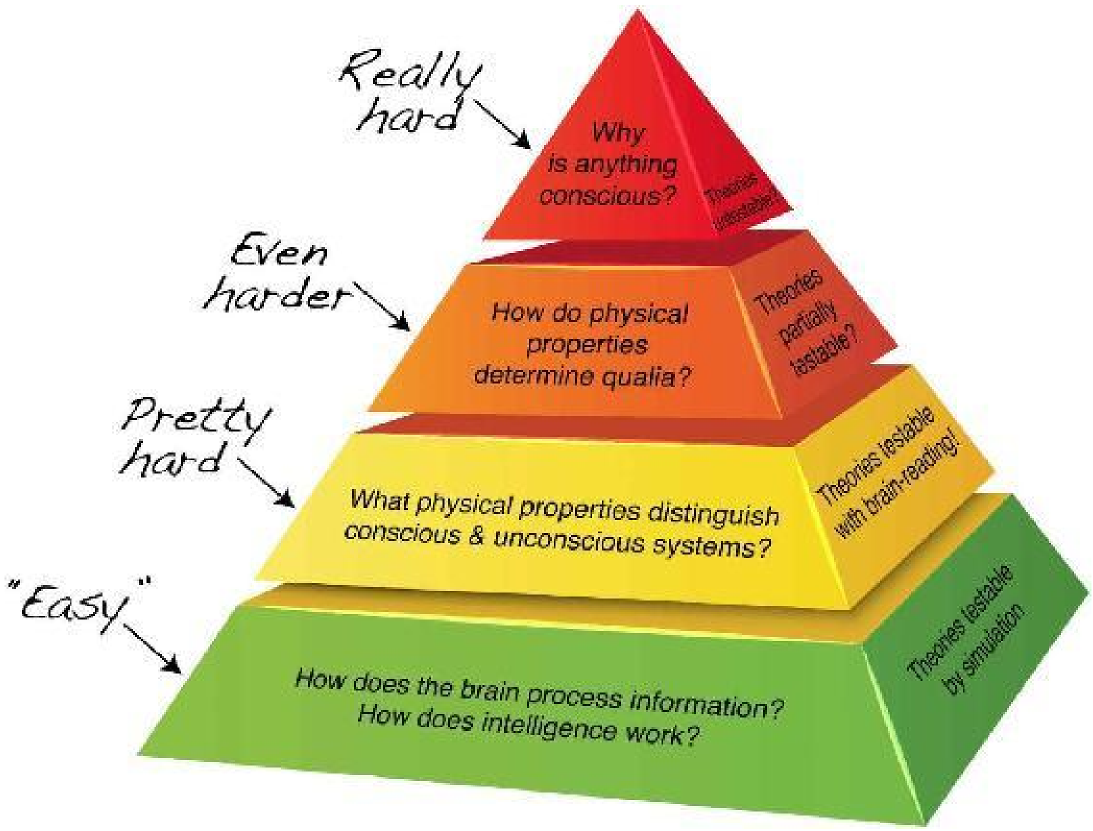
David'in vurguladığı gibi, gerçekten zihnin iki ayrı gizemi vardır. Birincisi, bir beynin bilgiyi nasıl işlediğinin gizemi var, David bunu "kolay" problemler olarak adlandırıyor. Örneğin, beyin duyusal girdiye nasıl bakar, onu nasıl yorumlar ve yanıt verir? Dil kullanarak iç durumu hakkında nasıl rapor verebilir? Bu sorular aslında son derece zor olsa da, tanımlarımız gereği bilinç gizemleri değil zekanın gizemleri: beynin nasıl hatırladığını, hesapladığını ve öğrendiğini soruyorlar. Dahası, kitabın ilk bölümünde yapay zeka araştırmacılarının, Go oynamaktan araba sürmeye, görüntüleri analiz etmeye ve doğal dili işlemeye kadar makinelerle bu "kolay sorunların" çoğunu çözme konusunda nasıl ciddi ilerleme kaydetmeye başladığını gördük.

Sonra, neden öznel bir deneyime sahip olduğunuza dair ayrı bir gizem var. *zor* sorun. Araba sürerken renkler, sesler, duygular ve bir benlik hissi yaşarsınız. Ama neden hiç bir şey yaşıyorsun? Kendi kendine giden bir araba herhangi bir şey yaşar mı? Kendi kendine giden bir arabaya karşı yarışıyorsanız, hem sensörlerden bilgi giriyor, hem işliyor hem de motor komutları veriyorsunuz. Ama öznel olarak *deneyimleme*

sürüş mantıksal olarak ayrı bir şeydir — isteğe bağlı mı ve öyleyse, buna ne sebep olur?

Bu zorlu bilinç sorununa fizik açısından yaklaşıyorum. Benim bakış açıma göre, bilinçli bir insan basitçe yeniden düzenlenmiş besindir. Öyleyse neden bir düzenleme bilinçli, diğeri değil? Dahası, fizik bize yiyeceğin sadece belirli bir şekilde düzenlenmiş çok sayıda kuark ve elektron olduğunu öğretir.

Peki hangi parçacık düzenlemeleri bilinçlidir ve hangileri değildir? * 1



Şekil 8.1: Zihni anlamak, bir problemler hiyerarşisini içerir. David Chalmers'ın "kolay" problemler dediği şey, öznel deneyimden bahsetmeden ortaya konulabilir. Fiziksel sistemlerin tamamının olmasa da bazılarının bilinçli olduğu aşikar gerçek, üç ayrı soruyu ortaya çıkarır. "Oldukça zor problemi" tanımlayan soruyu cevaplamak için bir teorimiz varsa, deneysel olarak test edilebilir. İşe yararsa, yukarıdaki daha zor soruların üstesinden gelmek için üzerine inşa edebiliriz.

Bu fizik perspektifinden hoşlandığım şey, biz insanlar olarak binlerce yıldır mücadele ettiğimiz zorlu sorunu, bilim yöntemleriyle daha kolay başa çıkılabilen daha odaklı bir versiyona dönüştürmesidir. Zorla başlamak yerine *sorun* neden bir parçacık düzenlemesinin bilinçli hissedebildiğini anlamak için zor bir şeyle başlayalım. *gerçek* bazı parçacık düzenlemeleri bilinçli hissederken diğerleri hissetmez. Örneğin, beyninizi oluşturan parçacıkların şu anda bilinçli bir düzenlemede olduğunu biliyorsunuz, ancak derin rüyasız uykudayken değil.

Bu fizik perspektifi, aşağıda gösterildiği gibi, bilinç hakkında üç ayrı zor soruya yol açar. şekil 8.1 . Her şeyden önce, parçacığın hangi özellikleri

düzenleme fark yaratır mı? Spesifik olarak, hangi fiziksel özellikler bilinçli ve bilinçsiz sistemleri birbirinden ayırır? Buna cevap verebilirsek, hangi AI sistemlerinin bilinçli olduğunu bulabiliriz. Daha yakın bir gelecekte, acil servis doktorlarının hangi yanıt vermeyen hastaların bilinçli olduğunu belirlemelerine de yardımcı olabilir.

İkinci olarak, fiziksel özellikler deneyimin neye benzediğini nasıl belirler? Özellikle ne belirler *qualia,* Gülün kırmızılığı, zil sesi, biftek kokusu gibi bilincin temel yapı taşları

mandalina tadı mı yoksa iğne batması acısı mı? * 2

Üçüncüsü, herhangi bir şey neden bilinçlidir? Başka bir deyişle, madde yığınlarının neden bilinçli olabileceğine dair derin ve keşfedilmemiş bir açıklama var mı, yoksa bu sadece dünyanın işleyişi hakkında açıklanamaz kaba bir gerçek mi?

Eski bir MIT meslektaşım olan bilgisayar bilimcisi Scott Aaronson, David Chalmers gibi, ilk soruyu gönülsüzce "oldukça zor sorun" (PHP) olarak adlandırdı. Bu ruhla, diğer ikisine "daha da zor sorun" diyelim (EHP) ve "gerçekten zor problem" (RHP), şekil 8.1 . * 3

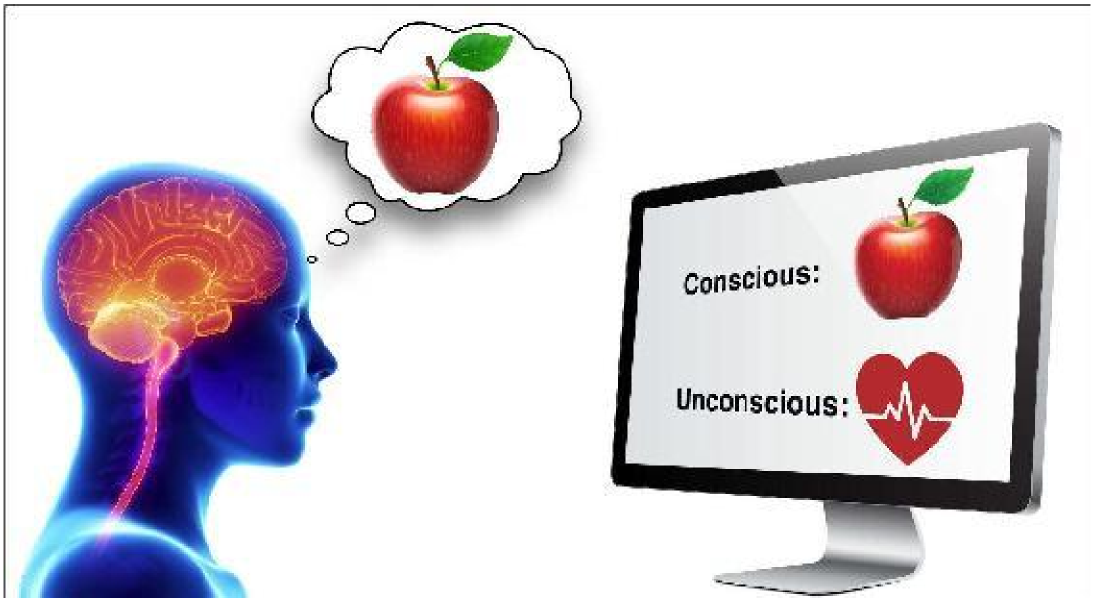
Bilinç Bilimin Ötesinde mi?

İnsanlar bana bilinç araştırmasının umutsuz bir zaman kaybı olduğunu söylediklerinde, verdikleri ana argüman bunun “bilim dışı” olduğu ve her zaman olacağıdır. Ama bu gerçekten doğrumu? Etkili Avusturyalı-İngiliz filozof Karl Popper, artık yaygın olarak kabul gören atasözü popüler hale getirdi "Eğer yanlışlanamazsa, bilimsel değildir." Başka bir deyişle, bilim tamamen teorileri gözlemlere karşı test etmekle ilgilidir: eğer bir teori prensipte bile test edilemiyorsa, o halde onu tahrif etmek mantıksal olarak imkansızdır, ki Popper'in tanımına göre bilim dışı olduğu anlamına gelir.

Öyleyse, aşağıdaki üç bilinç sorusundan herhangi birini yanıtlayan bilimsel bir teori olabilir mi? şekil 8.1 ? Lütfen sizi, cevabın yankılanan bir EVET! Olduğuna ikna etmeye çalışmama izin verin, en azından oldukça zor problem için: "Hangi fiziksel özellikler bilinçli ve bilinçsiz sistemleri birbirinden ayırır?" Herhangi bir fiziksel sistem verildiğinde, sistemin bilinçli olup olmadığı sorusuna "evet", "hayır" veya "emin değil" şeklinde yanıt veren bir teorisi olduğunu varsayalım. Beyninizi, beyninizin farklı bölümlerindeki bazı bilgi işlemlerini ölçen bir cihaza bağlayalım ve bu bilgiyi, bu bilginin hangi bölümlerinin bilinçli olduğunu tahmin etmek için bilinç teorisini kullanan ve size sunan bir bilgisayar programına aktaralım. bir ekranda gerçek zamanlı olarak tahminleri, olduğu gibi şekil 8.2 . Önce bir elma düşünürsünüz. Ekran, beyninizde farkında olduğunuz bir elma hakkında bilgi olduğunu, ancak beyin sapınızda nabzınız hakkında farkında olmadığınız bilgiler olduğunu size bildirir. Etkilenir misin? Teorinin ilk iki tahmini doğru olsa da, biraz daha titiz testler yapmaya karar veriyorsunuz. Annenizi düşünüyorsunuz ve bilgisayar size beyninizde anneniz hakkında bilgi olduğunu ama bunun farkında olmadığınızı söylüyor. Teori yanlış bir öngörüde bulundu, bu da onun reddedildiği ve Aristoteles mekaniği, ışıltılı eter, jeosentrik kozmoloji ve sayısız başka başarısız fikirle birlikte bilimsel tarihin çöplüklerine girdiği anlamına geliyor. Kilit nokta şudur: Teori yanlış olsa da, *ilmi!* Bilimsel olmasaydı, onu test edip göz ardı edemezdiniz.

Birisi bu sonucu eleştirebilir ve şunu söyleyebilir: *onlar* Neyin bilincinde olduğunuza dair hiçbir kanıtınız yok, hatta bilinçli olduğunuza dair hiçbir kanıtınız yok: Bilinçli olduğunuzu söylediğinizi duysalar bile, bilinçsiz bir zombi muhtemelen

aynı şeyi söyle. Ancak bu, bu bilinç teorisini bilim dışı yapmaz çünkü sizinle yerleri değiş tokuş edebilir ve doğru bir şekilde öngörüp öngörmediğini test edebilirler.
onların kendi bilinçli deneyimler.



Şekil 8.2: Bir bilgisayarın beyninizde işlenmekte olan bilgiyi ölçtüğünü ve bir bilinç teorisine göre onun hangi bölümlerinden haberdar olduğunuzu tahmin ettiğini varsayalım. Bu teoriyi, tahminlerinin doğru olup olmadığını, öznel deneyimlerinize uygun olup olmadığını kontrol ederek bilimsel olarak test edebilirsiniz.

Öte yandan, teori herhangi bir öngörüde bulunmayı reddederse, yalnızca sorgulandığında "emin değil" cevabını verirse, o zaman test edilemez ve dolayısıyla bilimsel değildir. Bu, sadece bazı durumlarda uygulanabilir olduğu için olabilir, çünkü gerekli hesaplamaları pratikte yapmak çok zordur veya beyin sensörleri işe yaramaz. Günümüzün en popüler bilimsel teorileri ortada bir yerde olma eğilimindedir ve tüm sorularımıza değil bazılarına test edilebilir yanıtlar verir. Örneğin, temel fizik teorimiz aynı anda aşırı derecede küçük (kuantum mekaniği gerektiren) ve aşırı derecede ağır (genel görelilik gerektiren) sistemler hakkındaki soruları yanıtlamayı reddedecektir, çünkü bu durumda hangi matematiksel denklemlerin kullanılacağını henüz bulamadık. . Bu çekirdek teori ayrıca tüm olası atomların tam kütlelerini tahmin etmeyi reddedecektir.

- bu durumda, gerekli denklemlere sahip olduğumuzu düşünüyoruz, ancak çözümlerini doğru bir şekilde hesaplamayı başaramadık. Bir teori boynunu uzatarak ve test edilebilir tahminler yaparak ne kadar tehlikeli bir şekilde yaşarsa, o kadar yararlı olur ve onu tüm öldürme girişimlerimizden sağ çıkarsa onu daha ciddiye alırız. Evet, sadece test edebiliriz *biraz* bilinç teorilerinin tahminleri, ama bu böyle

için *herşey* fiziksel teoriler. Neyi test edemeyeceğimiz konusunda mızmızlanmakla vakit kaybetmeyelim, ancak test etmeye başlayalım. *Yapabilmek* Ölçek!

Özetle, hangi fiziksel sistemlerin bilinçli olduğunu tahmin eden herhangi bir teori (oldukça zor problem), beyin süreçlerinizden hangilerinin bilinçli olduğunu tahmin edebildiği sürece bilimseldir. Bununla birlikte, test edilebilirlik sorunu, daha üst düzey sorular için daha az netleşir. şekil 8.1 . Bir teorinin kırmızı rengi öznel olarak nasıl deneyimlediğinizi tahmin etmesi ne anlama gelir? Ve eğer bir teori ilk etapta bilinç gibi bir şeyin neden olduğunu açıklamak isterse, onu deneysel olarak nasıl test edersiniz? Sırf bu soruların zor olması, onlardan kaçınmamız gerektiği anlamına gelmez ve aşağıda onlara geri döneceğiz. Ancak ilgili cevaplanmamış birkaç soruyla karşılaşıldığında, en kolay olanı ilk önce ele almanın akıllıca olduğunu düşünüyorum. Bu nedenle, MIT'deki bilinç araştırmam, doğrudan piramidin tabanına odaklanmıştır. şekil 8.1 . Kısa bir süre önce bu stratejiyi Princeton'dan fizikçi arkadaşım Piet Hut ile tartıştım; piramidin tepesini tabandan önce inşa etmeye çalışmanın, bize izin veren matematiksel temel olan Schrödinger denklemini keşfetmeden önce kuantum mekaniğinin yorumu hakkında endişelenmeye benzeyeceğini söyleyen şaka yaptı. deneylerimizin sonuçlarını tahmin edin.

Bilimin ötesinde olanı tartışırken, cevabın zamana bağlı olduğunu unutmamak önemlidir! Dört yüzyıl önce Galileo Galilei, matematiğe dayalı fizik teorilerinden o kadar etkilenmişti ki, doğayı “matematik dilinde yazılmış bir kitap” olarak tanımladı. Bir üzüm ve bir fındık fırlatırsa, yörüngelerinin şekillerini ve ne zaman yere düşeceklerini doğru bir şekilde tahmin edebilirdi. Yine de birinin neden yeşil, diğerinin kahverengi ya da birinin yumuşak ve diğerinin neden sert olduğuna dair hiçbir fikri yoktu - dünyanın bu yönleri o zamanlar bilimin ulaşamayacağı bir yerdeydi. Ama sonsuza kadar değil! James Clerk Maxwell 1861'de kendi adını taşıyan denklemlerini keşfettiğinde, ışık ve renklerin matematiksel olarak da anlaşılabileceği ortaya çıktı. 1925'te keşfedilen yukarıda bahsedilen Schrödinger denkleminin maddenin tüm özelliklerini tahmin etmek için kullanılabileceğini artık biliyoruz. yumuşak veya sert olanlar dahil. Teorik ilerleme her zamankinden daha fazla bilimsel öngörü sağlarken, teknolojik ilerleme daha da fazla deneysel testlere olanak sağladı: şu anda teleskoplar, mikroskoplar veya parçacık çarpıştırıcılarla incelediğimiz neredeyse her şey bir zamanlar bilimin ötesindeydi. Başka bir deyişle, bilimin kapsamı, Galileo'nun günlerinden bu yana, tüm fenomenlerin çok küçük bir kısmından, atom altı parçacıklar, kara delikler ve 13,8 milyar yıl önceki kozmik kökenlerimiz dahil olmak üzere büyük bir yüzdeye kadar dramatik bir şekilde genişledi. Bu şu soruyu gündeme getiriyor: Ne kaldı? Galileo'nun günlerinden bu yana bilimin kapsamı, tüm fenomenlerin çok küçük bir kısmından, atom altı parçacıklar, kara delikler ve 13,8 milyar yıl önceki kozmik kökenlerimiz dahil olmak üzere büyük bir yüzdeye kadar dramatik bir şekilde genişledi. Bu şu soruyu gündeme getiriyor: Ne kaldı? Galileo'nun günlerinden bu yana bilimin kapsamı, tüm fenomenlerin çok küçük bir kısmından, atom altı parçacıklar, kara delikler ve 13,8 milyar yıl önceki kozmik kökenlerimiz dahil olmak üzere büyük bir yüzdeye kadar dramatik bir şekilde genişledi. Bu şu soruyu gündeme getiriyor: Ne kaldı?

Bana göre, odadaki fil bilinçtir. Sadece bilinçli olduğunu bilmiyorsun, aynı zamanda *herşey* tam bir kesinlikle bilirsiniz - René Descartes'ın Galileo'nun zamanında işaret ettiği gibi, diğer her şey çıkarımdır. Teorik ve teknolojik ilerleme, sonunda bilinci bile kesin bir şekilde bilim alanına sokacak mı? Bilmiyoruz, tıpkı Galileo'nun bilmediği gibi

Bir gün ışığı ve maddeyi anlayıp anlayamayacağımızı. * 4 Tek bir şey garanti: Denemezsek başaramayız! Bu yüzden ben ve dünyadaki diğer birçok bilim insanı, bilinç teorilerini formüle etmek ve test etmek için çok çabalıyoruz.

Bilinçle İlgili Deneysel İpuçları

Şu anda kafamızda birçok bilgi işlem gerçekleşiyor. Hangisi bilinçli hangisi değil? Bilinç teorilerini ve ne tahmin ettiklerini keşfetmeden önce, geleneksel düşük teknolojili veya teknolojisiz gözlemlerden son teknoloji beyin ölçümlerine kadar deneylerin bize şimdiye kadar ne öğrettiğine bakalım.

Hangi Davranışlar Bilinçlidir?

Kafanızda 32 ile 17'yi çarparsanız, hesaplamanızın birçok iç işleyişinin farkında olursunuz. Ama farz edin ki ben size Albert Einstein'ın bir portresini gösteriyorum ve konunun adını söylemenizi söylüyorum. Bölümde gördüğümüz gibi
2, bu da bir hesaplama görevidir: beyniniz, girdisi çok sayıda piksel rengi hakkında gözlerinizden gelen bilgiler olan ve çıktısı ağzınızı ve ses tellerinizi kontrol eden kaslara bilgi olan bir işlevi değerlendiriyor. Bilgisayar bilimcileri bu görevi "görüntü sınıflandırması" ve ardından "konuşma sentezi" olarak adlandırıyor. Bu hesaplama, çarpma görevinizden çok daha karmaşık olsa da, bunu çok daha hızlı, görünüşte çaba harcamadan ve ayrıntıların farkında olmadan yapabilirsiniz. *Nasıl* sen yap. Öznel deneyiminiz yalnızca resme bakmaktan, bir tanınma duygusu yaşamaktan ve kendinizin "Einstein" dediğini duymaktan ibarettir.

Psikologlar, göz kırpma reflekslerinden nefes almaya, uzanmaya, tutmaya ve dengenizi korumaya kadar çok çeşitli başka görevleri ve davranışları da bilinçsizce gerçekleştirebileceğinizi uzun zamandır biliyorlar. Tipik olarak, ne yaptığınızın bilinçli olarak farkındasınız ama nasıl yaptığınızın değil. Öte yandan, alışılmadık durumlar, özdenetim, karmaşık mantıksal kurallar, soyut akıl yürütme veya dilin manipülasyonunu içeren davranışlar bilinçli olma eğilimindedir. Olarak bilinirler
bilincin davranışsal ilişkileri, ve psikologların "Sistem" olarak adlandırdığı çabalı, yavaş ve kontrollü düşünme biçimiyle yakından bağlantılıdırlar.
2. ” [5]

Yürüme, yüzme, bisiklete binme, araba kullanma, yazı yazma, tıraş olma, ayakkabı bağlama, bilgisayar oyunları ve piyano gibi kapsamlı uygulamalarla birçok rutini bilinçsizden bilinçdışına dönüştürebileceğiniz de bilinmektedir.
oynuyor. [6] Gerçekten de, uzmanların uzmanlıklarını en iyi şekilde "akış" durumunda olduklarında, yalnızca daha yüksek bir seviyede olanların farkında olduklarında ve bunu nasıl yaptıklarının düşük seviyeli ayrıntılarının farkında olmadan yaptıkları iyi bilinmektedir. Örneğin, okumayı ilk öğrendiğinizde olduğu gibi, her harfin farkında olarak sonraki cümleyi okumayı deneyin. Sadece sözcükler veya fikirler düzeyinde metnin bilincinde olduğunuz zamana kıyasla ne kadar yavaş olduğunu hissedebiliyor musunuz?

Gerçekte, bilinçsiz bilgi işleme sadece mümkün görünmekle kalmaz, aynı zamanda istisnadan çok kuraldır. Kanıt gösteriyor ki,

kabaca 10 [7] Her saniye duyu organlarımızdan beynimize giren bilgi parçacıklarıyla, yalnızca çok küçük bir kısmın farkında olabiliriz
10 ila 50 bit. [7] Bu, bilinçli olarak farkında olduğumuz bilgi işlemenin sadece buzdağının görünen kısmı olduğunu gösteriyor.

Birlikte ele alındığında, bu ipuçları bazı araştırmacıların, bilinçli bilgi işlemenin zihnimizin CEO'su olarak düşünülmesi gerektiğini, yalnızca karmaşık analiz gerektiren en önemli kararlarla ilgilenilmesi gerektiğini önermelerine yol açtı.
beynin her yerinden veriler. [8] Bu, neden tıpkı bir şirketin CEO'su gibi, çalışanlarının yaptıkları her şeyi bilerek dikkatinin dağılmasını istemediğini, ancak istenirse onları bulabileceğini açıklar. Bu seçici dikkati iş başında deneyimlemek için, "istenen" kelimesine tekrar bakın: bakışınızı "i" nin üzerindeki noktaya sabitleyin ve gözlerinizi hareket ettirmeden dikkatinizi noktadan tüm harfe ve sonra bütüne kaydırın. kelime. Retinanızdaki bilgiler aynı kalsa da, bilinçli deneyiminiz değişti. CEO metaforu, uzmanlığın neden bilinçsiz hale geldiğini de açıklıyor: CEO, okumayı ve yazmayı titizlikle çözdükten sonra, bu rutin görevleri yeni üst düzey zorluklara odaklanabilmek için bilinçsiz astlara devrediyor.

Bilinç Nerede?

Zekice yapılan deneyler ve analizler, bilincin yalnızca belirli davranışlarla değil, beynin belirli bölümleriyle de sınırlı olduğunu ileri sürdü. Baş şüpheliler hangileri? İlk ipuçlarının çoğu beyin lezyonu olan hastalardan geldi: kazalar, felçler, tümörler veya enfeksiyonların neden olduğu lokalize beyin hasarı. Ancak bu genellikle sonuçsuz kaldı. Örneğin, beynin arkasındaki lezyonların körlüğe neden olabileceği gerçeği, buranın görsel bilincin olduğu anlamına mı geliyor, yoksa sadece görsel bilginin, daha sonra bilinçli olacağı yere giderken oradan geçtiği anlamına mı geliyor? önce gözlerden geçer?

Lezyonlar ve tıbbi müdahaleler bilinçli deneyimlerin yerini tam olarak belirlemese de, seçeneklerin daraltılmasına yardımcı oldular. Örneğin, elimde gerçekten orada olduğu gibi ağrı yaşasam da, ağrı deneyiminin başka bir yerde olması gerektiğini biliyorum, çünkü bir kez bir cerrah elime hiçbir şey yapmadan el ağrımı kapattı: sadece omzumdaki sinirleri uyuşturdu. Dahası, bazı ampüteler, sanki hiç yokmuş gibi hissedilen hayali bir ağrı yaşarlar. Başka bir örnek olarak, bir keresinde sadece sağ gözümle baktığımda görme alanımın bir kısmının eksik olduğunu fark ettim - bir doktor retinamın gevşediğini belirledi ve yeniden bağladı. Aksine, belirli beyin lezyonları olan hastalar *hemineglect,* görme alanlarının yarısından gelen bilgileri de kaçırdıklarında, ancak eksik olduğunun farkında bile olmadıklarında - örneğin, tabaklarının sol yarısındaki yiyeceği fark edip yemediklerinde. Sanki dünyalarının yarısı kadar bilinç kaybolmuş gibi. Peki bu hasarlı beyin bölgelerinin uzaysal deneyimi oluşturması mı gerekiyor yoksa tıpkı retinamın yaptığı gibi sadece bilinç bölgelerine mekansal bilgi mi veriyorlar?

Öncü ABD-Kanadalı beyin cerrahı Wilder Penfield, 1930'larda beyin cerrahisi hastalarının şu anda adı verilen şeydeki belirli beyin alanlarını elektriksel olarak uyardığında vücutlarının farklı bölgelerine dokunulduğunu bildirdi.

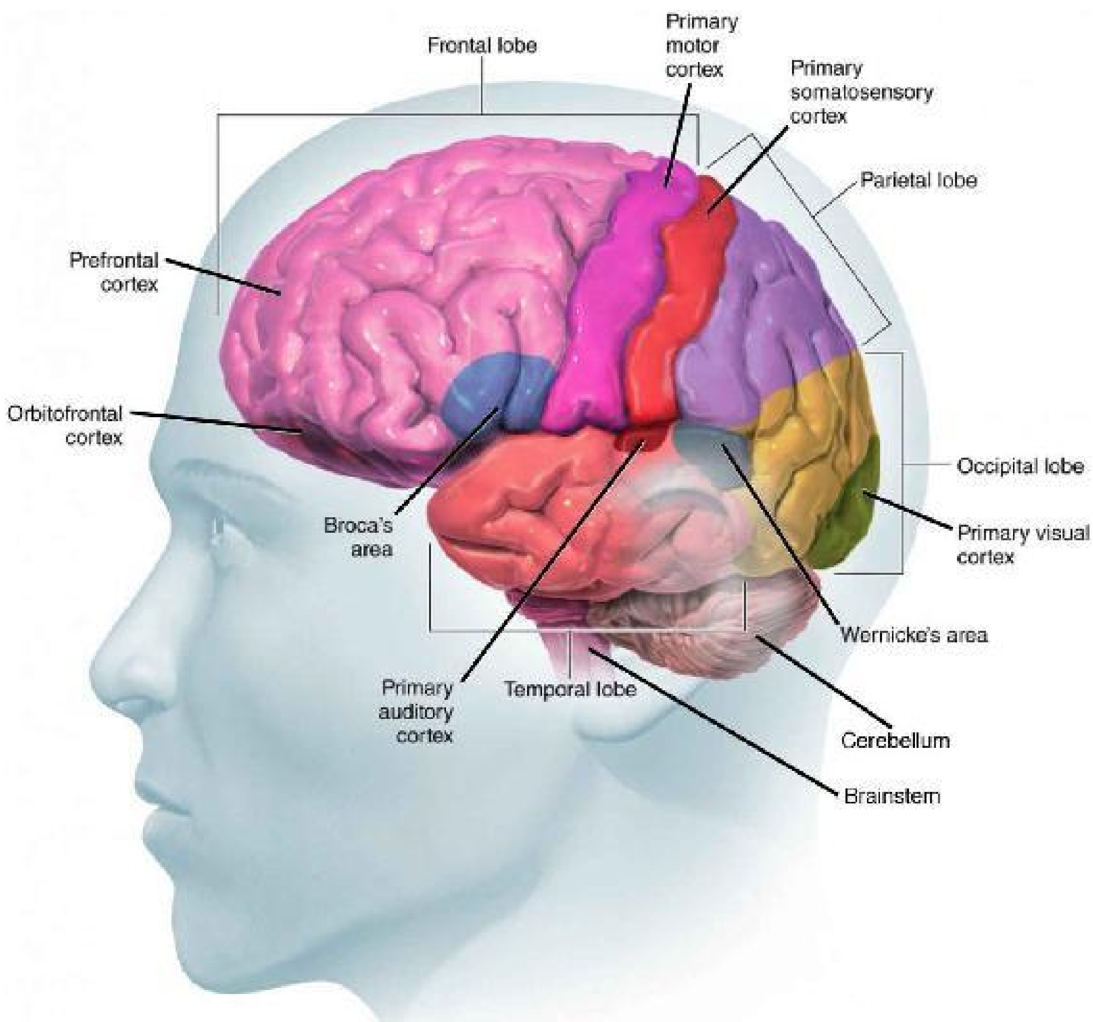
somatosensoriyel korteks (şekil 8.3). [9] Ayrıca, şu anda adı verilen yerde beyin bölgelerini uyardığında, vücutlarının farklı bölümlerini istemeden hareket ettirdiklerini de buldu. *motor korteks.* Ancak bu, beyin bölgelerindeki bilgi işlemenin dokunma ve hareket bilincine karşılık geldiği anlamına mı geliyor?

Neyse ki, modern teknoloji artık bize çok daha ayrıntılı ipuçları veriyor.

Yaklaşık yüz milyar nöronunuzun her bir ateşlemesini ölçemeyecek olsak da, beyin okuma teknolojisi fMRI, EEG, MEG, ECoG, ePhys ve floresan voltaj gibi göz korkutucu adlara sahip teknikleri içeren hızla ilerliyor. algılama. Fonksiyonel manyetik rezonans görüntüleme anlamına gelen fMRI, beyninizin yaklaşık saniyede bir milimetre ile 3 boyutlu bir haritasını çıkarmak için hidrojen çekirdeklerinin manyetik özelliklerini ölçer.

çözüm. EEG (elektroensefalografi) ve MEG (manyetoensefalografi) beyninizi saniyede binlerce kez haritalamak için başınızın dışındaki elektrik ve manyetik alanı ölçer, ancak zayıf çözünürlükle birkaç santimetreden daha küçük özellikleri ayırt edemez. Eğer titizseniz, bu üç tekniğin tamamen invazif olmadığını anlayacaksınız. Kafatasını açmanın bir sakıncası yoksa, ek seçenekleriniz var. ECoG (elektrokortikografi), beyninizin yüzeyine örneğin yüz tel yerleştirmeyi içerirken, ePhys (elektrofizyoloji), binlerce eşzamanlı konumdan voltajları kaydetmek için bazen bir insan saçından daha ince olan mikro telleri beynin derinliklerine sokmayı içerir. . Birçok epileptik hasta hastanede günler geçirirken, ECoG beyninin hangi bölümünün nöbetleri tetiklediğini ve rezeke edilmesi gerektiğini anlamak için kullanılır. ve bu arada sinirbilimcilerin bilinçli deneyler yapmasına izin vermeyi kabul ediyorum. Son olarak, flüoresan voltaj algılama, nöronların ateşleme sırasında ışık flaşları yayması için genetik olarak manipüle edilmesini içerir ve faaliyetlerinin bir mikroskopla ölçülmesini sağlar. Tüm tekniklerin dışında, en azından şeffaf beyinleri olan hayvanlarda, en fazla sayıda nöronu hızlı bir şekilde izleme potansiyeline sahiptir. *C. elegans* solucan 302 nöron ve larva zebra balığı yaklaşık

100.000.



Şekil 8.3: Görsel, işitsel, somatosensoriyel ve motor korteksler sırasıyla görme, işitme, dokunma hissi ve hareket aktivasyonu ile ilgilidir - ancak bu, nerede olduklarını kanıtlamaz *bilinç* görme, işitme, dokunma ve hareket oluşur. Nitekim, son araştırmalar birincil görsel korteksin beyincik ve beyin sapı ile birlikte tamamen bilinçsiz olduğunu öne sürüyor. Görüntü Lachina'nın izniyle (www.lachina.com).

Francis Crick, Christof Koch'u bilinç çalışması konusunda uyardı, ancak Christof pes etmeyi reddetti ve sonunda Francis'i kazandı. 1990'da, "bilincin sinirsel bağıntıları" (NCC'ler) dedikleri şey hakkında ufuk açıcı bir makale yazdılar ve hangi spesifik beyin süreçlerinin bilinçli deneyimlere karşılık geldiğini sordular. Binlerce yıldır düşünürler, beyinlerindeki bilgi işlemeye yalnızca öznel deneyimleriyle erişebiliyordu ve

davranış. Crick ve Koch, beyin okuma teknolojisinin birdenbire bu bilgiye bağımsız erişim sağladığına ve hangi bilgi işlemenin hangi bilinçli deneyime karşılık geldiğinin bilimsel çalışmasına izin verdiğine dikkat çekti. Yeterince emin, teknoloji odaklı ölçümler, NCC arayışını, nörobilimin oldukça yaygın bir parçası haline getirdi.

binlerce yayın en prestijli dergilere bile uzanıyor. [10]

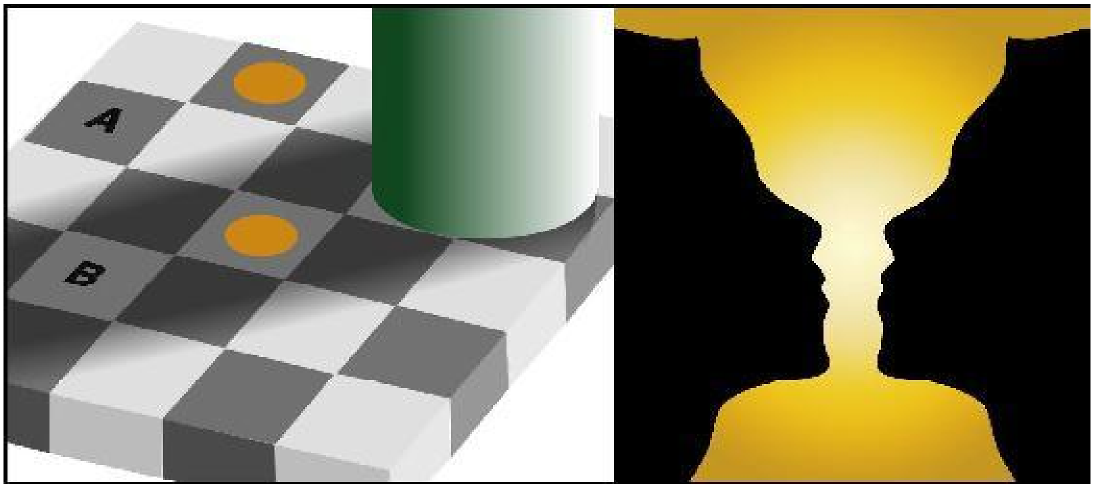
Şimdiye kadarki sonuçlar nelerdir? NCC dedektiflik çalışmasına bir lezzet katmak için, önce retinanızın bilinçli olup olmadığını veya yalnızca görsel bilgiyi kaydeden, işleyen ve bunu beyninizde öznel görsel deneyiminizin gerçekleştiği bir sisteme gönderen bir zombi sistemi olup olmadığını soralım. Sol panelinde şekil 8.4 , hangi kare daha koyu: A veya B etiketli kare? A, değil mi? Hayır, aslında aynı renkteler, parmaklarınız arasındaki küçük deliklerden onlara bakarak doğrulayabilirsiniz. Bu, görsel deneyiminizin tamamen retinanızda bulunmadığını kanıtlıyor, çünkü öyle olsaydı, aynı görüneceklerdi.

Şimdi sağ panele bakın şekil 8.4 . İki kadın mı yoksa bir vazo mu görüyorsun? Eğer

Yeterince uzun bakarsanız, retinanıza ulaşan bilgiler aynı kalsa bile, öznel olarak her ikisini de arka arkaya yaşarsınız. İki durumda beyninizde olanları ölçerek, farkı yaratan şeyi birbirinden ayırabilir - ve her iki durumda da aynı şekilde davranan retina değildir.

Bilinçli retina hipotezine yapılan ölüm darbesi, Christof Koch, Stanislas Dehaene ve işbirlikçilerinin öncülüğünü yaptığı "sürekli flaş bastırma" adı verilen bir teknikten gelir: Gözlerinizden birinin hızla değişen kalıpların karmaşık bir dizisini izlemesini sağlarsanız, o zaman keşfedilmiştir. Bu, görsel sisteminizin dikkatini dağıtacak ve tamamen farkında olmayacaksınız

diğer göze gösterilen hareketsiz bir görüntünün. [11] Özetle, retinanızda deneyimlemeden görsel bir görüntüye sahip olabilirsiniz ve (rüya görürken) bir görüntüyü retinanızda olmadan deneyimleyebilirsiniz. Bu, yüz milyondan fazla nöronu içeren karmaşık hesaplamalar yapsalar bile, iki retinanızın görsel bilincinizi bir video kameradan daha fazla barındırmadığını kanıtlıyor.



Şekil 8.4: Hangi kare daha koyu - A mı yoksa B mi? Sağda ne görüyorsunuz - vazo mu, iki kadın mı yoksa ikisi birden mi? Bu tür yanılsamalar, görsel bilincinizin gözünüzde veya görsel sisteminizin diğer erken aşamalarında olamayacağını kanıtlar, çünkü bu sadece resimdekine bağlı değildir.

NCC araştırmacıları ayrıca, beyin bölgelerinizden hangisini kesin olarak belirlemek için sürekli flaş bastırma, dengesiz görsel / işitsel yanılsamalar ve diğer hileler kullanır.
vardır bilinçli deneyimlerinizin her birinden sorumlu. Temel strateji, temelde her şeyin (duyusal girdiniz dahil) aynı olduğu iki durumda nöronlarınızın ne yaptığını karşılaştırmaktır - bilinçli deneyiminiz hariç. Beyninizin farklı davrandığı ölçülen bölümleri daha sonra NCC olarak tanımlanır.

Böyle bir NCC araştırması, yiyeceğinizi en iyi şekilde nasıl sindireceğinizi hesaplayan devasa yarım milyar nöronuyla enterik sinir sisteminizin yeri olmasına rağmen, bilincinizin hiçbirinin bağırsağınızda bulunmadığını kanıtlamıştır; Bunun yerine açlık ve mide bulantısı gibi duygular beyninizde üretilir. Benzer şekilde, bilincinizin hiçbiri beynin omuriliğe bağlanan ve nefes almayı, kalp atış hızını ve kan basıncını kontrol eden alt kısmı olan beyin sapında bulunmuyor gibi görünmüyor. Daha şok edici bir şekilde, bilinciniz beyincikinize kadar genişlemiyor gibi görünmüyor (şekil 8.3), tüm nöronlarınızın yaklaşık üçte ikisini içerir: beyincik tahrip olan hastalar, bir sarhoşluğu andıran gevrek konuşma ve beceriksiz hareketler yaşarlar, ancak tamamen bilinçli kalırlar.

Beyninizin hangi kısımlarının *vardır* bilinçten sorumlu

açık ve tartışmalı olmaya devam ediyor. Yakın zamanda yapılan bazı NCC araştırmaları, bilincinizin esas olarak talamusu (beyninizin ortasına yakın) ve korteksin arka kısmını (buruşuk altı katmanlı bir tabakadan oluşan dış beyin katmanı) içeren bir "sıcak bölgede" bulunduğunu ileri sürmektedir. eğer düzleştirilmişse,

büyük bir akşam yemeği peçetesinin alanı). [12] Aynı araştırma, tartışmalı bir şekilde, başın en arkasındaki birincil görsel korteksin, gözbebekleriniz ve retinalarınız kadar bilinçsiz olduğu için buna bir istisna olduğunu öne sürüyor.

Bilinç Ne Zaman?

Şimdiye kadar, ne tür bilgi işlemenin bilinçli olduğuna ve bilincin nerede oluştuğuna ilişkin deneysel ipuçlarına baktık. Fakat *ne zaman* oluyor mu Ben çocukken, kesinlikle gecikme veya gecikme olmaksızın olayların farkına vardığımızı düşünürdüm. Yine de öznel olarak bana öyle hissettirse de, beynimin duyu organlarımdan giren bilgileri işlemesi zaman aldığı için açıkça doğru olamaz. NCC araştırmacıları ne kadar süreyi dikkatli bir şekilde ölçtüler ve Christof Koch'un özeti, ışığın gözünüze girdiği andan itibaren yaklaşık çeyrek saniye sürdüğü.

siz onu bilinçli olarak görene kadar karmaşık nesne. [13] Bu, saatte elli beş mil hızla bir otobanda gidiyorsanız ve aniden önünüzde birkaç metre bir sincap görürseniz, bununla ilgili herhangi bir şey yapmak için çok geç demektir, çünkü zaten üzerinden geçtiniz. !

Özetle, Christof Koch dış dünyanın yaklaşık çeyrek saniye gerisinde kaldığını tahmin ederek, bilinçliliğiniz geçmişte yaşıyor. Şaşırtıcı bir şekilde, olaylara çoğu zaman farkına vardığınızdan daha hızlı tepki verebilirsiniz, bu da en hızlı tepkilerinizden sorumlu bilgi işlemenin bilinçsiz olması gerektiğini kanıtlar. Örneğin, gözünüze yabancı bir cisim yaklaşırsa, göz kırpma refleksiniz göz kapağınızı saniyenin onda biri kadar bir sürede kapatabilir. Sanki beyin sistemlerinizden biri görsel sistemden uğursuz bilgiler alıyor, gözünüzün çarpılma tehlikesi olduğunu hesaplıyor, göz kaslarınıza göz kırpma talimatlarını e-postayla gönderiyor ve aynı anda beyninizin bilinçli bölümüne e-posta ile "Hey, biz göz kırpacak. " Bu e-posta okunup bilinçli deneyiminize dahil edildiğinde,

Aslında, bu e-postayı okuyan sistem, vücudunuzun her yerinden gelen mesajlarla sürekli olarak bombardımana tutulur, bazıları diğerlerinden daha gecikir. Sinir sinyallerinin beyninize parmaklarınızdan ulaşması, uzaklığınız nedeniyle yüzünüzden daha uzun sürer ve görüntüleri analiz etmeniz seslerden daha uzun sürer çünkü daha karmaşıktır - bu nedenle Olimpiyat yarışları, görsel bir ipucu. Yine de burnunuza dokunursanız, bilinçli olarak burnunuzdaki ve parmak ucunuzdaki hissi aynı anda yaşarsınız ve ellerinizi çırparsanız, alkışları aynı şekilde görür, duyar ve hissedersiniz.

zaman. [14] Bu, bir olayla ilgili tam bilinçli deneyiminizin yaratılmadığı anlamına gelir.

son yavaş e-posta raporları alınıp analiz edilene kadar.

Fizyolog Benjamin Libet'in öncülüğünü yaptığı meşhur NCC deneyleri ailesi, bilinçsizce gerçekleştirebileceğiniz türden eylemlerin göz kırpma ve pinpon şutları gibi hızlı tepkilerle sınırlı olmadığını, aynı zamanda özgür iradeye atfedebileceğiniz belirli kararları da içerdiğini göstermiştir. —Beyin ölçümleri bazen kararınızı, siz gelmeden önce tahmin edebilir

başardığının bilincinde. [15]

Bilinç Teorileri

Henüz bilinci anlamasak da, onun çeşitli yönleri hakkında inanılmaz miktarda deneysel veriye sahip olduğumuzu gördük. Ama tüm bu veriler *beyin* öyleyse bize bilinçle ilgili herhangi bir şeyi nasıl öğretebilir?
makineler? Bu, mevcut deneysel alanımızın ötesinde büyük bir ekstrapolasyon gerektirir.
Başka bir deyişle, bir *teori.*

Neden Bir Teori?

Nedenini anlamak için, hadi bilinç teorilerini yerçekimi teorileriyle karşılaştıralım. Bilim adamları Newton'un yerçekimi teorisini ciddiye almaya başladılar çünkü içine koyduklarından daha fazlasını elde ettiler: bir peçeteye uyan basit denklemler, şimdiye kadar yapılmış her yerçekimi deneyinin sonucunu doğru bir şekilde tahmin edebilirdi. Bu nedenle, tahminlerini test edildiği alanın çok ötesinde ciddiye aldılar ve bu cesur ekstrapolasyonların milyonlarca ışık yılı genişliğindeki kümelerdeki galaksilerin hareketleri için bile işe yaradığı ortaya çıktı. Bununla birlikte, Merkür'ün Güneş etrafındaki hareketine ilişkin tahminler çok küçük bir oranda yanlıştı. Bilim adamları daha sonra Einstein'ın gelişmiş yerçekimi teorisini, genel göreliliği ciddiye almaya başladılar, çünkü tartışmalı olarak daha zarif ve ekonomikti ve Newton'un teorisinin yanlış gittiğini bile doğru bir şekilde tahmin ettiler.

Benzer şekilde, denklemleri bir peçeteye uyan matematiksel bir bilinç teorisi, beyinler üzerinde gerçekleştirdiğimiz tüm deneylerin sonuçlarını başarılı bir şekilde tahmin edebiliyorsa, o zaman sadece teorinin kendisini değil, aynı zamanda beyinlerin ötesinde bilinç için tahminlerini de ciddiye almaya başlarız. örneğin makinelerde.

Fizik Perspektifinden Bilinç

Bazı bilinç teorileri antik çağlara kadar uzanmasına rağmen, modern teorilerin çoğu nöropsikoloji ve sinirbilime dayanmaktadır ve açıklamaya çalışmaktadır.

ve beyinde meydana gelen sinirsel olaylar açısından bilinci tahmin etmek. [16]

Bu teoriler, bilincin sinirsel bağıntıları için bazı başarılı tahminler yapmış olsalar da, makine bilinci hakkında tahminlerde bulunamazlar. Beyinden makinelere sıçrama yapmak için, NCC'lerden PCC'lere genelleme yapmamız gerekir: *bilincin fiziksel bağlantıları,*

bilinçli hareket eden parçacıkların kalıpları olarak tanımlanır. Çünkü bir teori, yalnızca temel parçacıklar ve kuvvet alanları gibi fiziksel yapı taşlarına atıfta bulunarak neyin bilinçli olduğunu ve neyin olmadığını doğru bir şekilde tahmin edebiliyorsa, yalnızca beyinler için değil, aynı zamanda gelecekteki AI sistemleri de dahil olmak üzere diğer herhangi bir madde düzenlemesi için de tahminlerde bulunabilir. . Öyleyse bir fizik perspektifi ele alalım: Hangi parçacık düzenlemeleri bilinçlidir?

Ama bu gerçekten başka bir soruyu gündeme getiriyor: Bilinç kadar karmaşık bir şey, parçacıklar kadar basit bir şeyden nasıl yapılabilir? Sanırım bunun nedeni, parçacıklarının özelliklerinin üstünde ve ötesinde özelliklere sahip bir fenomen. İçinde

fizik, bu tür fenomenlere "ortaya çıkan" diyoruz. [17] Bunu bilinçten daha basit olan ortaya çıkan bir fenomene bakarak anlayalım: ıslaklık.

Bir damla su ıslaktır, ancak bir buz kristali ve bir buhar bulutu, aynı su moleküllerinden yapılmış olsalar bile değil. Neden? Çünkü ıslaklık özelliği sadece moleküllerin dizilişine bağlıdır. Tek bir su molekülünün ıslak olduğunu söylemek kesinlikle mantıklı değildir, çünkü ıslaklık olgusu yalnızca sıvı dediğimiz düzende düzenlenmiş birçok molekül olduğunda ortaya çıkar. Yani katılar, sıvılar ve gazların hepsi ortaya çıkan fenomenlerdir: Parçalarının toplamından daha fazlasıdır, çünkü parçacıklarının özelliklerinin üstünde ve ötesinde özelliklere sahiptirler. Parçacıklarının sahip olmadığı özelliklere sahiptirler.

Şimdi tıpkı katılar, sıvılar ve gazlar gibi, bilincin de parçacıklarının özelliklerinin üzerinde ve ötesinde özelliklerle ortaya çıkan bir fenomen olduğunu düşünüyorum. Örneğin, derin uykuya girmek, yalnızca parçacıkları yeniden düzenleyerek bilinci söndürür. Aynı şekilde donarak ölürsem bilincim kaybolur, bu da parçacıklarımı daha talihsiz bir şekilde yeniden düzenlerdi.

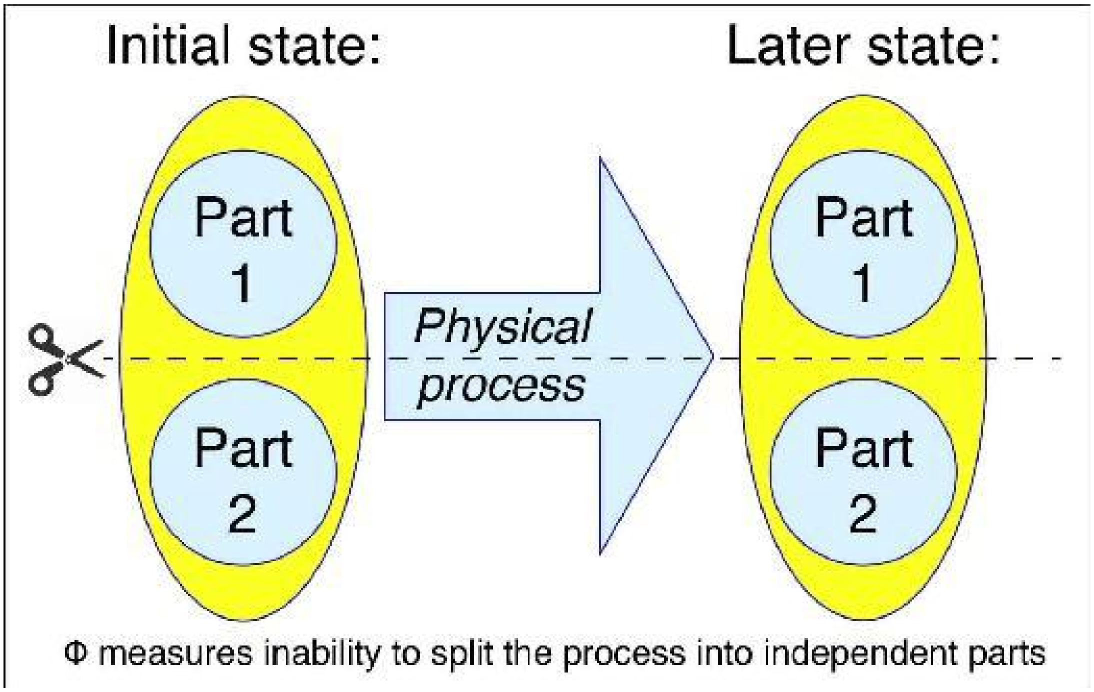
Sudan suya kadar bir şey yapmak için çok sayıda parçacığı bir araya getirdiğinizde

beyin, gözlemlenebilir özelliklere sahip yeni fenomenler ortaya çıkar. Biz fizikçiler, genellikle dışarı çıkıp ölçebileceğiniz küçük bir sayı dizisi ile tanımlanabilen bu ortaya çıkan özellikleri incelemeyi seviyoruz - maddenin ne kadar viskoz olduğu, ne kadar sıkıştırılabilir olduğu vb. Miktarlar. Örneğin, bir madde katı olacak kadar viskozsa, ona katı diyoruz, aksi takdirde sıvı olarak adlandırıyoruz. Ve eğer bir sıvı sıkıştırılamazsa, buna sıvı diyoruz, aksi takdirde elektriği ne kadar iyi ilettiğine bağlı olarak ona gaz veya plazma diyoruz.

Bilgi Olarak Bilinç

Öyleyse bilinci ölçen benzer nicelikler olabilir mi? İtalyan sinirbilimci Giulio Tononi böyle bir miktar önerdi ve

"Entegre bilgi" Yunan harfi ile gösterilir Φ (*Phi),* Temel olarak bir sistemin farklı bölümlerinin birbirini ne kadar bildiğini ölçer (bkz.
şekil 8.5).



Şekil 8.5: Zaman geçtikçe bir sistemin başlangıç durumunu yeni bir duruma dönüştüren fiziksel bir süreç verildiğinde, *entegre bilgi* Φ süreci bağımsız parçalara bölmedeki yetersizliği ölçer. Her parçanın gelecek durumu, diğer parçanın ne yaptığına değil, yalnızca kendi geçmişine bağlıysa, o zaman Φ = 0: Bir sistem dediğimiz şey aslında birbiriyle hiç iletişim kurmayan iki bağımsız sistemdir.

Giulio ile ilk olarak, onu ve Christof Koch'u davet ettiğim Porto Riko'daki bir 2014 fizik konferansında tanıştım ve o, Galileo ve Leonardo da Vinci ile doğrudan kaynaşmış nihai rönesans adamı olarak beni vurdu. Sessiz tavrı, inanılmaz sanat, edebiyat ve felsefe bilgisini gizleyemezdi ve mutfak şöhreti ondan önce geldi: Kozmopolit bir televizyon muhabiri, geçenlerde bana Giulio'nun birkaç dakika içinde en lezzetli salatayı nasıl hazırladığını anlatmıştı d hayatında tattı. Çok geçmeden onun yumuşak sözlü tavrının arkasında, kurumun önyargıları ve tabularına bakmaksızın, kanıtları onu götürdüğü her yerde takip edecek korkusuz bir zeka olduğunu anladım. Tıpkı Galileo'nun, yermerkezciliğe meydan okumama yönündeki kuruluş baskısına rağmen matematiksel hareket teorisini takip ettiği gibi,

entegre bilgi teorisi (IIT).

On yıllardır bilincin, bilginin belirli karmaşık şekillerde işlendiğinde hissettiği yol olduğunu tartışıyordum. [18] HTE buna katılıyor ve belirsiz “belirli karmaşık yollar” ifademi kesin bir tanımla değiştiriyor: bilgi işlemenin entegre edilmesi gerekiyor, yani Φ büyük olmalı. Giulio'nun bunun için argümanı basit olduğu kadar güçlüdür: bilinçli sistemin birleşik bir bütün halinde bütünleştirilmesi gerekir, çünkü eğer onun yerine iki bağımsız parçadan oluşsaydı, o zaman bir yerine iki ayrı bilinçli varlık gibi hissederlerdi. Başka bir deyişle, bir beynin veya bilgisayarın bilinçli bir parçası diğerleriyle iletişim kuramıyorsa, geri kalanı onun öznel deneyiminin bir parçası olamaz.

Giulio ve çalışma arkadaşları beynin manyetik uyarıma tepkisini ölçmek için EEG kullanarak Φ'nin basitleştirilmiş bir versiyonunu ölçtüler. “Bilinç detektörü” gerçekten iyi çalışıyor: Hastaların uyandıklarında veya rüya gördüklerinde bilinçli olduklarını, anestezi aldıklarında veya derin uykuda bilinçsiz olduklarını belirledi. Hatta "kilitli" sendromundan muzdarip, normal şekilde hareket edemeyen veya iletişim kuramayan iki hastada bilinci bile keşfetti. [19] Dolayısıyla bu, gelecekte doktorlar için belirli hastaların bilinçli olup olmadığını anlamaları için umut verici bir teknoloji olarak ortaya çıkıyor.

Fizikte Bilincin Demirlenmesi

IIT yalnızca sınırlı sayıda durumda olabilen ayrık sistemler için tanımlanır, örneğin bir bilgisayar belleğindeki bitler veya açık veya kapalı olabilen aşırı basitleştirilmiş nöronlar. Bu maalesef, HTE'nin sürekli olarak değişebilen geleneksel fiziksel sistemlerin çoğu için tanımlanmadığı anlamına gelir - örneğin, bir parçacığın konumu veya bir manyetik alanın kuvveti herhangi bir

sonsuz sayıda değer. [20] IIT formülünü bu tür sistemlere uygulamaya çalışırsanız, genellikle Φ sonsuz olan yararsız bir sonuç alırsınız. Kuantum mekanik sistemler ayrık olabilir, ancak orijinal IIT kuantum sistemleri için tanımlanmamıştır. Öyleyse HTE ve diğer bilgi temelli bilinç teorilerini sağlam bir fiziksel temele nasıl bağlayabiliriz?

Bunu, madde yığınlarının bilgiyle ilgili yeni ortaya çıkan özelliklere sahip olabileceği konusunda 2. bölümde öğrendiklerimizi temel alarak yapabiliriz. Bir şeyin bilgiyi depolayabilen bir hafıza cihazı olarak kullanılabilmesi için birçok uzun ömürlü duruma sahip olması gerektiğini gördük. Biz de gördük *computronium,* Hesaplamalar yapabilen bir madde, ayrıca karmaşık dinamikler gerektirir: fizik yasalarının, keyfi bilgi işlemeyi uygulayabilecek kadar karmaşık şekillerde değiştirmesi gerekir. Son olarak, örneğin bir sinir ağının öğrenme için nasıl güçlü bir alt tabaka olduğunu gördük, çünkü basitçe fizik yasalarına uyarak, istenen hesaplamaları daha iyi ve daha iyi bir şekilde uygulamak için kendini yeniden düzenleyebilir. Şimdi ek bir soru soruyoruz: Bir madde damlasını öznel bir deneyime sahip kılan nedir? Başka bir deyişle, hangi koşullar altında bir madde damlası bu dört şeyi yapabilir?

1. hatırla
2. hesaplama
3. öğrenmek
4. deneyim

2. bölümde ilk üçünü araştırdık ve şimdi dördüncü ile uğraşıyoruz. Tıpkı Margolus ve Toffoli'nin bu terimi icat ettiği gibi *bilgisayar* rastgele hesaplamalar yapabilen bir madde için, terimini kullanmayı seviyorum *sentronyum* çoğu için

öznel deneyime sahip (duyarlı) genel madde. * 5

Fakat bilinç aslında fiziksel bir fenomense, nasıl bu kadar fiziksel olmayan hissedebilir? Fiziksel alt tabakasından bu kadar bağımsız nasıl hissediyor? Sanırım bunun nedeni *dır-dir* fiziksel alt tabakasından, içinde bir model olduğu şeyden oldukça bağımsızdır! Bölüm 2'de dalgalar, anılar ve hesaplamalar dahil olmak üzere birçok güzel substrat bağımsız model örneğiyle karşılaştık. Nasıl sadece kendi parçalarından (ortaya çıkan) fazlası olmadıklarını, daha ziyade kendi parçalarından bağımsız olduklarını, kendi hayatlarını sürdürdüklerini gördük. Örneğin, gelecekte simüle edilmiş bir zihin veya bilgisayar oyunu karakterinin, alt tabakadan bağımsız olacağı için Windows, Mac OS, Android telefon veya başka bir işletim sisteminde çalışıp çalışmadığını bilmesinin hiçbir yolu olmadığını gördük. Bilgisayarının mantık kapılarının transistörlerden, optik devrelerden veya diğer donanımlardan yapılıp yapılmadığını da söyleyemezdi. Ya da fiziğin temel yasaları nedir - evrensel bilgisayarların inşasına izin verdikleri sürece herhangi bir şey olabilirler.

Özetle, bilincin fiziksel olmayan bir fenomen olduğunu düşünüyorum çünkü dalgalar ve hesaplamalara benziyor: belirli fiziksel alt tabakasından bağımsız özelliklere sahip. Bu mantıksal olarak bilgi olarak bilinç fikrinden kaynaklanır. Bu, gerçekten sevdiğim radikal bir fikre yol açar: Bilinç, bilginin belirli şekillerde işlendiğinde hissettiği yolsa, o zaman substrattan bağımsız olmalıdır; önemli olan bilgi işlemenin yapısıdır, bilgi işlemeyi yapan konunun yapısı değildir. Diğer bir deyişle, bilinç iki kattan bağımsızdır!

Gördüğümüz gibi, fizik uzay-zamanda hareket eden parçacıklara karşılık gelen kalıpları tanımlar. Parçacık düzenlemeleri belirli ilkelere uyuyorsa, parçacık alt tabakasından oldukça bağımsız olan ve tamamen farklı bir his veren ortaya çıkan fenomenlere yol açarlar. Bunun harika bir örneği, computronium'daki bilgi işlemedir. Ama şimdi bu fikri başka bir düzeye taşıdık: *Bilgi işlemenin kendisi belirli ilkelere uyuyorsa, bilinç dediğimiz daha yüksek düzeyde ortaya çıkan fenomeni ortaya çıkarabilir.*

Bu, bilinçli deneyiminizi maddeden bir değil iki seviye yukarıya yerleştirir. Zihninizin fiziksel olmadığını hissetmesine şaşmamalı!

Bu bir soruyu gündeme getiriyor: Bilinçli olmak için bilgi işlemenin uyması gereken bu ilkeler nelerdir? Koşulların ne olduğunu biliyormuş gibi yapmıyorum
yeterli bilinci garanti etmek için, ama işte dört *gerekli* araştırmamda araştırdığım ve üzerine bahis oynadığım koşullar:

Prensip	Tanım
Bilgi prensip	Bilinçli bir sistemin önemli bir bilgi depolama kapasitesi vardır.
Dinamikler prensip	Bilinçli bir sistemin önemli bilgi işleme kapasitesi vardır.
Bağımsızlık prensip	Bilinçli bir sistemin dünyanın geri kalanından önemli bir bağımsızlığı vardır.
Entegrasyon prensip	Bilinçli bir sistem neredeyse bağımsız parçalardan oluşamaz.

Dediğim gibi, bilginin belirli şekillerde işlenirken hissettirdiği yolun bilinç olduğunu düşünüyorum. Bu, bir sistemin bilinçli olmak için ilk iki ilkeyi ifade ederek bilgiyi depolayabilmesi ve işleyebilmesi gerektiği anlamına gelir. Hafızanın uzun sürmesine gerek olmadığını unutmayın: Anılarına rağmen tamamen bilinçli görünen Clive Wearing'in bu dokunaklı videosunu izlemenizi tavsiye ederim.

bir dakikadan az sürer. [21] Bence bilinçli bir sistem aynı zamanda dünyanın geri kalanından oldukça bağımsız olmalıdır, çünkü aksi takdirde öznel olarak herhangi bir bağımsız varoluşa sahip olduğunu hissetmezdi. Son olarak, Giulio Tononi'nin savunduğu gibi, bilinçli sistemin birleşik bir bütün halinde bütünleştirilmesi gerektiğini düşünüyorum, çünkü eğer iki bağımsız parçadan oluşsaydı, o zaman bir yerine iki ayrı bilinçli varlık gibi hissedeceklerdi. İlk üç ilke şu anlama gelir: *özerklik:* Sistemin bilgileri dışarıdan çok fazla müdahale olmaksızın tutup işleyebilmesi ve dolayısıyla kendi geleceğini belirlemesi. Dört ilkenin tümü birlikte bir sistemin otonom olduğu ancak parçalarının olmadığı anlamına gelir.

Bu dört ilke doğruysa, işimiz bizim için biçilmiş kaftan: onları somutlaştıran matematiksel olarak titiz teorileri aramalı ve deneysel olarak test etmeliyiz. Ek ilkelere ihtiyaç olup olmadığını da belirlememiz gerekiyor. HTE'nin doğru olup olmadığına bakılmaksızın, araştırmacılar rakip teoriler geliştirmeye çalışmalı ve mevcut tüm teorileri daha iyi deneylerle test etmelidir.

Bilinç Tartışmaları

Bilinç araştırmasının bilim dışı bir saçmalık ve anlamsız bir zaman kaybı olup olmadığı konusundaki sürekli tartışmayı zaten tartışmıştık. Ek olarak, bilinç araştırmalarının en ileri noktasında son zamanlarda tartışmalar var - en aydınlatıcı bulduklarımı inceleyelim.

Giulio Tononi'nin HTE'si son zamanlarda sadece övgü değil, bazıları aşağılayıcı eleştiriler de aldı. Scott Aaronson son zamanlarda blogunda şunları söylemişti: "Bana göre, Bütünleşik Bilgi Teorisinin yanlış olduğu gerçeği - açıkça yanlıştır, özüne giden nedenlerden dolayı - onu tüm matematiksel teorilerin ilk% 2'si gibi bir şeye koyar. bilinç önerdi. Bana öyle geliyor ki, neredeyse tüm rakip bilinç teorileri çok belirsiz, kabarıktı

ve sadece yanlışlığı arzulayabilecekleri şekilde şekillendirilebilir. " [22] Hem Scott hem de Giulio'nun itibarına, New York Üniversitesi'ndeki bir atölye çalışmasında IIT'yi tartışmalarını izlediğimde asla darbe alamadılar ve birbirlerinin argümanlarını kibarca dinlediler. Aaronson, bazı basit mantık kapısı ağlarının son derece yüksek entegre bilgiye (Φ) sahip olduğunu gösterdi ve açıkça bilinçli olmadıkları için HTE'nin yanlış olduğunu savundu. Giulio, eğer inşa edilirlerse,

olur bilinçli olun ve Scott'ın tersine varsayımı insan merkezli olarak önyargılıydı, sanki bir mezbaha sahibi hayvanların sırf konuşamadıkları ve insanlardan çok farklı oldukları için bilinçli olamayacaklarını iddia ediyormuş gibi. İkisinin de hemfikir olduğu analizim, entegrasyonun yalnızca bir *gerekli* bilinç durumu (Scott'ın iyi olduğu) veya ayrıca *yeterli* durum (Giulio'nun iddia ettiği). İkincisi, açıkça daha güçlü ve daha tartışmalı bir iddiadır,

Umarım yakında deneysel olarak test edebiliriz. [23]

Tartışmalı bir diğer HTE iddiası, bugünün bilgisayar mimarilerinin bilinçli olamayacağıdır, çünkü mantık kapılarının bağlanma şekli çok düşüktür.

entegrasyon. [24] Başka bir deyişle, kendinizi nöronlarınızın ve sinapslarınızın her birini doğru bir şekilde simüle eden gelecekteki yüksek güçlü bir robota yüklerseniz, bu dijital klon sizden ayırt edilemez şekilde görünse, konuşsa ve hareket etse bile Giulio bunun bilinçsiz olacağını iddia ediyor. öznel deneyime sahip olmayan zombi - kendinizi bir göreve yüklerseniz hayal kırıklığı yaratır

öznel ölümsüzlük için. * [6] Bu iddia, hem David Chalmers hem de yapay zeka profesörü Murray Shanahan tarafından, beyninizdeki sinir devrelerini yavaş yavaş varsayımsal olarak değiştirirseniz ne olacağını hayal ederek meydan okudu.
mükemmel bir şekilde simüle eden dijital donanım. [25] Senin olmasına rağmen *davranış* Simülasyon mükemmel bir varsayımla yapıldığından, değişimden etkilenmeyecektir. *deneyim* Giulio'ya göre, başlangıçta bilinçten sonunda bilinçdışına geçecekti. Ama her zamankinden daha fazla değiştirildiği gibi, arada nasıl hissederdi? Beyninizin, görme alanınızın üst yarısının bilinçli deneyiminden sorumlu olan kısımları değiştirildiğinde, görsel manzaranızın aniden kaybolduğunu, ancak gizemli bir şekilde

"kör gören" hastalar tarafından bildirildiği gibi, yine de ne olduğunu biliyor muydunuz? [26]
Bu derinden rahatsız edici olur, çünkü herhangi bir farkı bilinçli olarak deneyimleyebilirseniz, sorulduğunda arkadaşlarınıza da bundan bahsedebilirsiniz - ancak varsayımla, davranışınız değişemez. Varsayımlarla uyumlu tek mantıksal olasılık, herhangi bir şeyin bilincinizden kaybolduğu tam olarak aynı anda, zihninizin gizemli bir şekilde değiştirilerek ya yalan söylemeye ve deneyiminizin değiştiğini inkar etmenize ya da olayların değiştiğini unutmanıza neden olmasıdır. farklı.

Öte yandan, Murray Shanahan, aynı kademeli değiştirme eleştirisinin aynı seviyeye getirilebileceğini kabul ediyor. *hiç* Bilinçli olmadan bilinçli hareket edebileceğinizi iddia eden bir teori, bu nedenle hareket etme ve bilinçli olmanın aynı şey olduğu ve bu nedenle önemli olan tek şeyin dışarıdan gözlemlenebilir davranış olduğu sonucuna varmak isteyebilirsiniz. Ama o zaman, daha iyi bilseniz bile, rüya görürken bilinçsiz olduğunuzu tahmin etme tuzağına düşersiniz.

Üçüncü bir HTE tartışması, bilinçli bir varlığın ayrı ayrı bilinçli parçalardan oluşup oluşmayacağıdır. Örneğin, içindeki insanlar kendilerini kaybetmeden bir bütün olarak toplum bilinç kazanabilir mi? Bilinçli bir beynin de kendi başına bilinçli parçaları olabilir mi? HTE'den gelen tahmin kesin bir “hayır” dır, ancak herkes ikna olmuş değil. Örneğin, beyinlerinin iki yarısı arasındaki iletişimi ciddi şekilde azaltan lezyonları olan bazı hastalar, sağ beyinlerinin, hastaların neden olmadıklarını veya anlamadıklarını iddia ettikleri şeyleri sol ellerine yaptırdığı "uzaylı el sendromu" yaşarlar. diğer ellerini "yabancı" ellerini dizginlemek için kullandıklarını. Kafalarında iki ayrı bilincin olmadığından nasıl bu kadar emin olabiliriz, biri sağ yarıkürede konuşamıyor, diğeri ise

sol hemisfer bu tüm konuşmayı yapıyor ve ikisi adına konuştuğunu iddia ediyor? İki insan beyni arasında doğrudan bir iletişim bağlantısı kurmak için geleceğin teknolojisini kullandığınızı ve bu bağlantının kapasitesini, beyinler arasındaki iletişim içlerinde olduğu kadar verimli olana kadar kademeli olarak artırdığınızı hayal edin. İki bireysel bilincin birdenbire ortadan kaybolduğu ve HTE'nin öngördüğü gibi tek bir birleşik bilinciyle değiştirildiği bir an gelir miydi, yoksa geçiş kademeli mi olur, böylece ortak bir deneyim ortaya çıkmaya başlasa bile bireysel bilinçler bir biçimde bir arada var olur mu?

Bir başka ilginç tartışma da, deneylerin ne kadar bilinçli olduğumuzu hafife alıp almadığı. Daha önce gördük ki *hissetmek* renkler, şekiller, nesneler ve görünüşte önümüzde olan her şeyi içeren çok büyük miktarda bilginin görsel olarak bilincindeyiz, deneyler,

sadece bunun çok küçük bir kısmını hatırlayın ve bildirin. [27] Bazı araştırmacılar, bu tutarsızlığı, bazen "erişime sahip olmayan bilince" sahip olup olamayacağımızı, yani,

daha sonra kullanmak için çalışma belleğimize sığmayacak kadar karmaşık. [28] Örneğin, deneyimlediğinizde *dikkatsiz körlük* Düz bakışta bir nesneyi fark edemeyecek kadar dikkatinizin dağılması, bu, bilinçli bir görsel deneyiminiz olmadığı anlamına gelmez.
sadece sizin çalışma belleğinizde depolanmamış olması. [29] Körlükten çok unutkanlık olarak mı sayılmalı? Diğer araştırmacılar, insanlara deneyimlediklerini söyledikleri şey hakkında güvenilemeyeceği fikrini reddediyor ve bunun sonuçları konusunda uyardı. Murray Shanahan, hastaların yeni bir harika ilaç sayesinde ağrının tamamen azaldığını bildirdikleri, ancak yine de bir hükümet heyeti tarafından reddedilen bir klinik çalışma hayal ediyor: “Hastalar sadece acı çekmediklerini düşünüyor.
Sinirbilim sayesinde daha iyisini biliyoruz. " [30] Öte yandan, ameliyat sırasında yanlışlıkla uyanan hastalara çileyi unutturacak bir ilaç verildiği durumlar olmuştur. Sonraki raporlarına güvenmeli miyiz?
acı çekmedin mi? [31]

AI Bilinci Nasıl Hissedebilir?

Gelecekteki bir yapay zeka sistemi bilinçliyse, öznel olarak ne deneyimleyecek? Bu, bilincin "daha da zor problemi" nin özüdür ve bizi burada tasvir edilen ikinci zorluk seviyesine zorlar. şekil 8.1 . Şu anda bu soruyu yanıtlayan bir teoriden yoksun değiliz, aynı zamanda tam olarak cevaplamanın mantıksal olarak mümkün olup olmadığından bile emin değiliz. Sonuçta, tatmin edici bir cevap kulağa nasıl gelebilir? Kör doğmuş birine kırmızı rengin neye benzediğini nasıl açıklarsınız?

Neyse ki, tam bir cevap verememememiz, kısmi cevaplar vermemizi engellemiyor. İnsan duyu sistemini inceleyen zeki uzaylılar, muhtemelen renklerin iki boyutlu bir yüzeydeki (görsel alanımız) her noktayla ilişkili hissettiren nitelikler olduğunu, seslerin mekansal olarak yerelleştirilmiş gibi hissetmediğini ve ağrıların ise vücudumuzun farklı bölümleri. Retinalarımızın üç tür ışığa duyarlı koni hücresine sahip olduğunu keşfetmekten, üç ana rengi deneyimlediğimizi ve diğer tüm renk kalitelerinin bunların birleştirilmesinden kaynaklandığı sonucuna varabilirler. Nöronların beyne bilgi aktarmasının ne kadar sürdüğünü ölçerek, saniyede ondan fazla bilinçli düşünce veya algı deneyimlemediğimiz sonucuna varabilirler. ve televizyonumuzda saniyede yirmi dört kare hızla film izlediğimizde, bunu bir durağan görüntü dizisi olarak değil, sürekli hareket olarak deneyimliyoruz. Adrenalinin kan dolaşımımıza ne kadar hızlı salındığını ve parçalanmadan önce ne kadar kaldığını ölçerek, saniyeler içinde başlayan ve dakikalarca süren öfke patlamaları hissettiğimizi tahmin edebilirler.

Fizik temelli benzer argümanları uygulayarak, yapay bir bilincin nasıl hissedebileceğine dair bazı eğitimli tahminlerde bulunabiliriz. Her şeyden önce, olası AI deneyimlerinin alanı *Kocaman* biz insanların deneyimleyebilecekleriyle karşılaştırıldığında. Duyularımızın her biri için bir sınıf kalitemiz var, ancak AI'lar çok daha fazla sensör türüne ve dahili bilgi temsillerine sahip olabilir, bu nedenle bir AI olmanın zorunlu olarak bir insan olmaya benzediğini varsaymanın tuzağından kaçınmalıyız.

İkincisi, beyin büyüklüğünde bir yapay bilinç saniyede bizden milyonlarca kat daha fazla deneyime sahip olabilir, çünkü elektromanyetik sinyaller ışık hızında - nöron sinyallerinden milyonlarca kat daha hızlı hareket eder. Ancak, daha büyük

Bölüm 4'te gördüğümüz gibi, AI küresel düşünceleri ne kadar yavaş olursa, bilgi zamanının tüm parçaları arasında akmasına izin vermelidir. Bu nedenle Dünya ölçeğinde bir "Gaia" AI'nın saniyede yalnızca on bilinçli deneyime sahip olmasını bekleriz. , tıpkı bir insan gibi ve galaksi büyüklüğündeki bir YZ, her 100.000 yılda bir, yalnızca bir küresel düşünceye sahip olabilir - yani şimdiye kadar Evrenimizin tüm tarihi boyunca yaklaşık yüz deneyimden fazlası olamaz! Bu, büyük yapay zekalara, hesaplamaları bunları idare edebilen en küçük alt sistemlere devretmek, işleri hızlandırmak için, tıpkı bilinçli zihnimizin göz kırpma refleksini küçük, hızlı ve bilinçsiz bir alt sisteme devretmesi gibi, görünüşte karşı konulamaz bir teşvik verecektir. Yukarıda beynimizdeki bilinçli bilgi işlemenin, aksi takdirde bilinçsiz bir buzdağının görünen kısmı gibi göründüğünü görmemize rağmen, Gelecekteki yapay zekalar için durumun daha da aşırı olmasını beklemeliyiz: eğer tek bir bilince sahiplerse, o zaman içinde yer alan neredeyse tüm bilgi işlemenin farkında olmayacaktır. Dahası, zevk aldığı bilinçli deneyimler son derece karmaşık olsa da, daha küçük parçalarının hızlı faaliyetlerine kıyasla salyangoz tempolu.

Bu, bilinçli bir varlığın parçalarının da bilinçli olup olamayacağına dair yukarıda bahsedilen tartışmayı gerçekten gündeme getiriyor. IIT tahmin etmemektedir; bu, gelecekte astronomik olarak büyük bir YZ bilinçliyse, neredeyse tüm bilgi işlemenin bilinçsiz olduğu anlamına gelir. Bu, daha küçük AI'lardan oluşan bir uygarlığın iletişim yeteneklerini tek bir bilinçli kovan zihninin ortaya çıktığı noktaya kadar geliştirmesi durumunda, çok daha hızlı bireysel bilinçlerinin aniden söndüğü anlamına gelir. Öte yandan, HTE öngörüsü yanlışsa, kovan zihni daha küçük bilinçli zihinlerin panopili ile bir arada var olabilir. Gerçekte, mikroskobik seviyeden kozmiye kadar tüm seviyelerde iç içe geçmiş bir bilinç hiyerarşisi bile hayal edilebilir.

Yukarıda gördüğümüz gibi, insan beynimizdeki bilinçsiz bilgi işleme, zahmetsiz, hızlı ve otomatik düşünme biçimiyle bağlantılı görünmektedir.
psikologlar "Sistem 1" diyor. [32] Örneğin, Sistem 1, bilincinize, görsel girdi verilerinin oldukça karmaşık analizinin, hesaplamanın nasıl gerçekleştiğine dair size herhangi bir fikir vermeden en iyi arkadaşınızın geldiğini belirlediğini bildirebilir. Sistemler ve bilinç arasındaki bu bağın geçerli olduğu kanıtlanırsa, bu terminolojiyi yapay zekalara genellemek, bilinçsiz alt birimlere verilen tüm hızlı rutin görevleri YZ'nin Sistemi olarak ifade etmek cazip olacaktır 1. YZ'nin zahmetli, yavaş ve kontrollü küresel düşüncesi. eğer bilinçli olsaydı, AI'nın Sistemi 2 olurdu. Biz insanlar da benim “Sistem 0” olarak adlandıracağım şeyi içeren bilinçli deneyimler yaşarız: gerçekleşen ham pasif algı

hareket etmeden veya düşünmeden oturduğunuzda ve sadece etrafınızdaki dünyayı gözlemlediğinizde bile. Sistem 0, 1 ve 2 giderek daha karmaşık görünüyor, bu yüzden sadece ortadaki sistemin bilinçsiz görünmesi dikkat çekicidir. IIT bunu, Sistem 0'daki ham duyusal bilginin çok yüksek entegrasyonla ızgara benzeri beyin yapılarında depolandığını söyleyerek açıklarken, Sistem 2'nin şu anda farkında olduğunuz tüm bilgilerin sizi etkileyebileceği geri bildirim döngüleri nedeniyle yüksek entegrasyona sahip olduğunu söyleyerek açıklar. gelecekteki beyin durumları. Öte yandan, Scott Aaronson'ın yukarıda bahsedilen HTE eleştirisini tetikleyen tam da bilinçli ızgara tahminiydi. Özetle, eğer oldukça zor bir bilinç problemini çözen bir teori, bir gün zorlu deneysel testlerden geçebilir, böylece öngörülerini ciddiye almaya başlayabiliriz.

Öznel deneyimimizin bazı yönleri açıkça evrimsel kökenimize kadar uzanır, örneğin kendini koruma (yemek, içmek, öldürülmekten kaçınma) ve üremeyle ilgili duygusal arzularımız. Bu, açlık, susuzluk, korku veya cinsel istek gibi nitelikleri asla yaşamayan YZ yaratmanın mümkün olması gerektiği anlamına gelir. Son bölümde gördüğümüz gibi, oldukça zeki bir yapay zeka, neredeyse herhangi bir yeterince iddialı hedefe sahip olacak şekilde programlanmışsa, bu hedefe ulaşabilmek için muhtemelen kendini korumak için çaba gösterecektir. Bununla birlikte, bir yapay zeka topluluğunun parçası iseler, güçlü insan ölüm korkumuzdan yoksun olabilirler: kendilerini yedekledikleri sürece, kaybedecekleri tek şey, en son yedeklemelerinden bu yana biriktirdikleri anılardır. , yedeklenen yazılımlarının kullanılacağından emin oldukları sürece.

Yapay bir bilinç, özgür iradeye sahip olduğunu hisseder miydi? Filozoflar bin yıl boyunca, *Biz* özgür iradeye sahip olmak

Sorunun nasıl tanımlanacağı konusunda bile fikir birliğine varmadan, [33] Tartışılması daha kolay olan farklı bir soru soruyorum. Sizi, cevabın basitçe "Evet, herhangi bir bilinçli karar verici öznel olarak

hissetmek biyolojik ya da yapay olmasına bakılmaksızın özgür iradeye sahip olduğunu. " Kararlar, iki uç nokta arasında bir yelpazede yer alır:

1. Bu seçimi neden yaptığınızı tam olarak biliyorsunuz.
2. Neden bu seçimi yaptığınız hakkında hiçbir fikriniz yok - bir hevesle rastgele seçmişsiniz gibi hissettirdi.

Özgür irade tartışmaları genellikle hedefe yönelik karar verme davranışımızı fizik yasalarıyla uzlaştırmaya yönelik bir mücadelenin etrafında toplanır: Yaptığınız şey için aşağıdaki iki açıklama arasında seçim yapıyorsanız, o zaman hangisi doğrudur: *"Ona bir randevu sordum çünkü ondan gerçekten hoşlandım"* veya *"Parçacıklarım fizik kanunlarına göre hareket ederek bana bunu yaptırdı"?* Ama son bölümde gördük ki *her ikisi de* doğrudur: Hedefe yönelik davranış gibi hissettiren şey, hedefsiz deterministik fizik yasalarından ortaya çıkabilir. Daha spesifik olarak, bir sistem (beyin veya yapay zeka) tip 1 kararını verdiğinde, belirleyici bir algoritma kullanarak neye karar vereceğini hesaplar ve karar vermiş gibi hissetmesinin nedeni, aslında ne yapacağını hesaplarken karar vermiş olmasıdır. Dahası, vurguladığı gibi

Seth Lloyd, [34] Neredeyse tüm hesaplamalar için, sonuçlarını belirlemenin onları çalıştırmaktan daha hızlı bir yolu olmadığını söyleyen ünlü bir bilgisayar bilimi teoremi var. Bu, bir saniyeden daha kısa bir süre içinde ne yapmaya karar vereceğinizi anlamanızın genellikle imkansız olduğu anlamına gelir, bu da özgür iradeye sahip olma deneyiminizi güçlendirmeye yardımcı olur. Aksine, bir sistem (beyin veya yapay zeka) tip 2'ye karar verdiğinde, kararını rastgele sayı üreteci gibi davranan bir alt sistemin çıktısına dayandırmak için zihnini programlar. Beyinlerde ve bilgisayarlarda, etkili rastgele sayılar, gürültüyü artırarak kolayca üretilir. 1'den 2'ye kadar bir kararın nereye düştüğüne bakılmaksızın, hem biyolojik hem de yapay bilinçler özgür iradeye sahip olduklarını hissederler:

Bazıları bana nedenselliği aşağılayıcı bulduklarını, bunun düşünce süreçlerini anlamsız hale getirdiğini ve onları "salt" makineler haline getirdiğini söylüyor. Böyle olumsuzlukları saçma ve haksız buluyorum. Her şeyden önce, bildiğim kadarıyla, bilinen Evrenimizdeki en şaşırtıcı derecede sofistike fiziksel nesneler olan insan beyni hakkında "sadece" hiçbir şey yoktur. İkincisi, hangi alternatifi tercih ederlerdi? Kararlarını veren kendi düşünce süreçleri (beyinleri tarafından gerçekleştirilen hesaplamalar) olmasını istemiyorlar mı? Öznel özgür irade deneyimleri, hesaplamalarının içeriden nasıl hissettiğidir: bir hesaplamanın sonucunu, onu bitirene kadar bilmezler. Hesaplamanın *dır-dir* karar.

Anlam

Bu kitabın başlangıç noktasına dönerek bitirelim: Hayatın geleceğinin nasıl olmasını istiyoruz? Bir önceki bölümde, dünyanın dört bir yanındaki farklı kültürlerin nasıl olumlu deneyimlerle dolu bir gelecek aradıklarını gördük, ancak neyin olumlu sayılması gerektiği ve farklı yaşam formları için neyin iyi olduğu arasında nasıl değiş tokuş yapılacağı konusunda fikir birliği ararken, büyüleyici derecede çetrefilli tartışmalar ortaya çıkıyor. . Ancak bu tartışmaların bizi odadaki filden uzaklaştırmasına izin vermeyelim: Hiç deneyim yoksa, yani bilinç yoksa olumlu deneyimler olamaz. Diğer bir deyişle, bilinç olmadan mutluluk, iyilik, güzellik, anlam ya da amaç olamaz - sadece astronomik bir uzay israfı. *Bizim Evrenimiz bilinçli varlıklara anlam vermiyor, bilinçli varlıklar Evrenimize anlam veriyor.* Öyleyse gelecek için dilek listemizdeki ilk hedef, kozmosumuzdaki biyolojik ve / veya yapay bilincin yok olmasına neden olmak yerine muhafaza etmek (ve umarız ki genişletmek) olmalıdır.

Bu çabayı başarırsak, o zaman biz insanlar daha akıllı makinelerle bir arada var olma konusunda ne hissedeceğiz? Yapay zekanın görünüşte amansız yükselişi sizi rahatsız ediyor mu ve öyleyse neden? Bölüm 3'te, yapay zeka destekli teknolojinin, politik irade var olduğu sürece güvenlik ve gelir gibi temel ihtiyaçlarımızı karşılamasının nispeten kolay olması gerektiğini gördük. Ancak, belki de iyi beslenmenin, giyinmenin, barındırılmanın ve eğlenmenin yeterli olmadığından endişeleniyorsunuzdur. Yapay zekanın tüm pratik ihtiyaçlarımızı ve arzularımızı karşılayacağından emin olursak, yine de bakımlı hayvanat bahçesi hayvanları gibi hayatımızda anlam ve amaçtan yoksun olduğumuzu hissedebilir miyiz?

Geleneksel olarak, biz insanlar öz değerimizi genellikle *insan istisnacılığı:* Gezegendeki en zeki varlıklar olduğumuz ve bu nedenle benzersiz ve üstün olduğumuz inancı. AI'nın yükselişi bizi bundan vazgeçmeye ve daha alçakgönüllü olmaya zorlayacak. Ama belki de yine de yapmamız gereken bir şey bu: Sonuçta, başkalarına (bireyler, etnik gruplar, türler vb.) Karşı kibirli üstünlük kavramlarına sarılmak geçmişte korkunç sorunlara neden oldu ve emekliliğe hazır bir fikir olabilir. Gerçekten de, insan istisnası

Geçmişte sadece kedere neden olmakla kalmadı, aynı zamanda insanlığın gelişmesi için de gereksiz görünüyor: bilimde, sanatta ve önemsediğimiz her şeyde bizden çok daha gelişmiş, barışçıl bir dünya dışı uygarlığı keşfedersek, bu muhtemelen insanları hayatlarında anlam ve amaç yaşamaya devam ediyor. Ailelerimizi, arkadaşlarımızı ve daha geniş topluluklarımızı ve bize anlam ve amaç veren tüm faaliyetleri elimizde tutabilirdik, umarım kibirden başka bir şey kaybetmemiş olurduk.

Geleceğimizi planlarken, sadece kendi hayatlarımızın değil, Evrenimizin de anlamını ele alalım. Burada en sevdiğim fizikçilerden ikisi, Steven Weinberg ve Freeman Dyson, taban tabana zıt görüşleri temsil ediyor. Parçacık fiziğinin standart modeli üzerine yapılan temel çalışmalardan dolayı Nobel Ödülü'nü kazanan Weinberg, ünlü bir şekilde, "Evren ne kadar çok görünürse
anlaşılırsa, o kadar anlamsız da görünür. " [35] Dyson ise 6. bölümde gördüğümüz gibi çok daha iyimser: Evrenimizin *oldu* anlamsız, hayatın artık onu daha fazla anlamla doldurduğuna, hayatın kozmosa yayılmayı başarması durumunda gelebilecek en iyisi olduğuna inanıyor. 1979'daki çığır açan makalesini şöyle bitirdi: "Weinberg'in evreni mi yoksa benim mi

gerçeğe daha yakın mı? Çok geçmeden bir gün önce öğreneceğiz. " [36] Evrenimiz, Dünya yaşamının tükenmesine neden olduğumuz için veya bilinçsiz zombi AI'nın Evrenimizi ele geçirmesine izin verdiğimiz için kalıcı olarak bilinçsiz olmaya geri dönerse, Weinberg maçta haklı çıkacaktır.

Bu perspektiften, bu kitapta zekanın geleceğine odaklanmış olsak da, bilincin geleceği daha da önemli çünkü anlamı mümkün kılan şey bu. Filozoflar, bu ayrıma Latin gitmeyi severler. *keskinlik (* akıllıca düşünme yeteneği) ile

duyarlılık (qualia'yı öznel olarak deneyimleme yeteneği). Biz insanlar kimliğimizi olmak üzerine inşa ettik *Homo sapiens,* çevredeki en akıllı varlıklar. Daha akıllı makinelerin alçakgönüllülüğüne hazırlanırken, kendimizi şu şekilde yeniden markalandırmamızı
Homo sentiens!

ALT ÇİZGİ:

- "Bilincin" tartışmasız bir tanımı yoktur. Geniş ve insan merkezli olmayan tanımı kullanıyorum *bilinç = öznel deneyim.*

- YZ'lerin bu anlamda bilinçli olup olmadığı, YZ'nin yükselişinin ortaya çıkardığı en çetrefilli etik ve felsefi sorunlar için önemli olan şeydir: YZ'ler acı çekebilir mi? Hakları olmalı mı? Yüklemek öznel bir intihar mı? Yapay zeka ile dolu gelecekteki bir evren, nihai zombi kıyameti olabilir mi?

- Zekayı anlama sorunu, üç ayrı bilinç problemi ile birleştirilmemelidir: hangi fiziksel sistemlerin bilinçli olduğunu tahmin etmenin "oldukça zor problemi", kaliteyi tahmin etmenin "daha da zor problemi" ve " *Gerçekten mi* neden her şeyin bilinçli olduğuna dair zor bir problem.

- Bilincin "oldukça zor problemi" bilimseldir, çünkü hangi beyin süreçlerinizin bilinçli olduğunu tahmin eden bir teori deneysel olarak test edilebilir ve yanlışlanabilirken, bilimin iki zor problemi tam olarak nasıl çözebileceği şu anda belirsizdir.

- Sinirbilim deneyleri, birçok davranışın ve beyin bölgesinin bilinçsiz olduğunu ve bilinçli deneyimlerimizin çoğunun çok daha büyük miktarlarda bilinçsiz bilginin olgudan sonraki özetini temsil ettiğini göstermektedir.

- Bilinç tahminlerini beyinden makinelere genellemek bir teori gerektirir. Bilinç, belirli bir tür parçacığı veya alanı değil, oldukça özerk ve entegre olan belirli bir tür bilgi işlemeyi gerektiriyor gibi görünüyor, böylece tüm sistem oldukça özerktir, ancak parçaları değildir.

- Bilinç çok fiziksel olmayan hissedebilir çünkü iki kat bağımsızdır: eğer bilinç, bilginin belirli karmaşık şekillerde işlenirken hissettiği yolsa, o zaman önemli olan yalnızca bilgi işlemenin yapısıdır, konuyu yapan maddenin yapısı değil. bilgi işlem.

- Yapay bilinç mümkünse, olası YZ deneyimlerinin alanı, biz insanların deneyimleyebileceğine kıyasla çok büyük olabilir ve geniş bir nitelik ve zaman ölçeği yelpazesini kapsar - hepsi özgür iradeye sahip olma hissini paylaşır.

- Bilinç olmadan anlam olamayacağına göre, Evrenimizin bilinçli varlıklara anlam vermesi değil, bilinçli varlıklar Evrenimize anlam veriyor.

- Bu, biz insanlar her zamankinden daha akıllı makineler tarafından alçaltılmaya hazırlanırken, esas olarak varlığımızın rahatlığını *Homo sentiens,* değil *Homo sapiens.*

---

* 1 Alternatif bir bakış açısı *madde ikiliği* - canlı varlıklar cansız olanlardan farklı olduğu için

"anima", "élan vital" veya "ruh" gibi fiziksel olmayan bazı maddeler içerirler. Bilim adamları arasındaki madde ikiliği için destek giderek azaldı. Nedenini anlamak için vücudunuzun

yaklaşık 10'dan yapılmıştır 29 Söyleyebileceğimiz kadarıyla basit fizik kanunlarına göre hareket eden kuarklar ve elektronlar. Tüm parçacıklarınızı izleyebilecek gelecekteki bir teknoloji hayal edin: fizik yasalarına tam olarak uydukları tespit edilirse, o zaman sözde ruhunuzun parçacıklarınız üzerinde hiçbir etkisi olmaz, bu nedenle bilinçli zihninizin ve hareketlerinizi kontrol etme yeteneğinin hiçbir şeyi olmaz. bir ruhla yap. Parçacıklarınızın, ruhunuz tarafından itildikleri için bilinen fizik yasalarına uymadığı tespit edilirse, bu kuvvetlere neden olan yeni varlık, tanım gereği, tıpkı yeni alanlar üzerinde çalıştığımız gibi çalışabileceğimiz fiziksel bir varlık olacaktır. ve geçmişte yeni parçacıklar.

* 2 Sözlük tanımına göre "qualia" kelimesini, öznel deneyimin bireysel örneklerini ifade etmek için kullanıyorum - yani, deneyime neden olduğu iddia edilen herhangi bir maddeyi değil, öznel deneyimin kendisini kastediyorum. Bazı insanların bu sözcüğü farklı şekilde kullandığına dikkat edin.

* 3 Başlangıçta RHP'yi "çok zor sorun" olarak adlandırmıştım, ancak bu bölümü David Chalmers'a gösterdikten sonra, bana e-postayla, *"Gerçekten zor bir sorun"* onun gerçekten kastettiği ile eşleşecek şekilde: "İlk iki sorun (en azından bu şekilde söylersek) benim tasarladığım şekliyle gerçekten zor sorunun bir parçası olmadığından, üçüncü sorun ise, belki de" gerçekten zor "yerine Üçüncünün benim kullanımıma uyması "çok zor". "

* 4 Kitabımda keşfettiğim gibi, fiziksel gerçekliğimiz tamamen matematiksel ise (bilgiye dayalı, gevşek bir şekilde konuşan) *Matematiksel Evrenimiz,* o zaman gerçekliğin hiçbir yönü - bilinç bile - bilimin kapsamının ötesinde değildir. Aslında, bu bakış açısıyla, bilincin gerçekten zor olan problemi, matematiksel bir şeyin fiziksel olarak nasıl hissedilebileceğini anlamakla aynı problemdir: eğer matematiksel bir yapının bir parçası bilinçliyse, o zaman geri kalanını dış fiziksel dünya olarak deneyimleyecektir.

* 5 Daha önce "perceptronium" u sentronium ile eşanlamlı olarak kullanmış olsam da, bu isim çok dar bir tanım önermektedir, çünkü algılar yalnızca duyusal girdiye dayalı olarak algıladığımız öznel deneyimlerdir - örneğin, rüyalar ve dahili olarak üretilen düşünceler hariç.

* 6 Bu iddia ile bilincin substrattan bağımsız olduğu fikri arasında potansiyel bir gerilim vardır, çünkü en düşük seviyede bilgi işleme farklı olsa bile, davranışları belirlediği daha yüksek seviyelerde tanım gereği aynıdır.

Sonsöz

FLI Ekibinin Hikayesi

Şu anda hayatın en üzücü yönü, bilimin bilgiyi toplumun bilgelik topladığından daha hızlı toplamasıdır.

Isaac asimov

Sevgili okuyucum, zekanın, amaçların ve anlamın kökenini ve kaderini keşfettikten sonra kitabın sonunda buradayız. Peki bu fikirleri eyleme nasıl dönüştürebiliriz? Somut olarak ne yapmalıyız *yapmak* Geleceğimizi olabildiğince iyi hale getirmek için? 9 Ocak'ta San Francisco'dan Boston'a dönerken, şu anda burada pencere koltuğumda otururken kendime sorduğum soru tam olarak bu.
2017, Asilomar'da düzenlediğimiz AI konferansından, bu yüzden sizlerle düşüncelerimi paylaşarak bu kitabı bitireyim.

Meia, hazırlık ve organizasyonla geçen birçok kısa gecenin ardından yanımda uykusuna yetişiyor. Vay be, ne kadar vahşi bir hafta oldu! Elon Musk ve Larry Page gibi girişimciler ve akademi ve DeepMind, Google, Facebook gibi şirketlerden yapay zeka araştırma liderleri de dahil olmak üzere, bu kitapta bahsettiğim hemen hemen tüm insanları birkaç günlüğüne bu Porto Riko devam filminde bir araya getirmeyi başardık. Apple, IBM, Microsoft ve Baidu'nun yanı sıra ekonomistler, hukuk bilimcileri, filozoflar ve diğer harika düşünürler (bkz. şekil 9.1). Sonuçlar yüksek beklentilerimin bile yerini aldı ve hayatın geleceği konusunda uzun zamandır sahip olduğumdan daha iyimser hissediyorum. Bu sonsözde size nedenini anlatacağım.

FLI Doğdu

On dört yaşında nükleer silahlanma yarışını öğrendiğimden beri, teknolojimizin gücünün, onu yönettiğimiz bilgelikten daha hızlı büyüdüğünden endişe duyuyorum. Bu nedenle, bu meydan okumayla ilgili ilk kitabıma bir bölüm eklemeye karar verdim. *Matematiksel Evrenimiz,* geri kalanı öncelikle fizikle ilgili olsa bile. 2014 için bir Yeni Yıl kararı aldım, kişisel olarak ne yapabileceğimi ciddi bir şekilde düşünmeden hiçbir şeyden şikayet etmeme artık izin verilmedi ve o Ocak ayında kitap turum sırasında sözümü tuttum: Meia ve ben çok şey yaptık teknolojik idare yoluyla yaşamın geleceğini iyileştirmeye odaklanan bir tür kar amacı gütmeyen organizasyon başlatmak hakkında beyin fırtınası.

Ona "Doom & Gloom Institute" ve "Let's-End-End-the-Future Institute" dan olabildiğince farklı bir pozitif isim vermemiz konusunda ısrar etti. Future of Humanity Institute çoktan alındığından beri, daha kapsayıcı olma avantajına sahip olan Future of Life Institute (FLI) üzerinde bir araya geldik. 22 Ocak'ta kitap turu bizi Santa Cruz'a götürdü ve California Sun Pasifik üzerinde batarken, eski dostumuz Anthony Aguirre ile akşam yemeğinin tadını çıkardık ve onu bizimle güçlerimizi birleştirmesi için ikna ettik. O tanıdığım en bilge ve en idealist insanlardan biri değil, aynı zamanda başka bir kar amacı gütmeyen kuruluş olan *Temel Sorular Enstitüsü*

(görmek http://fqxi.org), on yıldan fazla bir süredir benimle.

Ertesi hafta tur beni Londra'ya götürdü. Yapay zekanın geleceği aklımda olduğundan, beni nezaketle DeepMind'ın merkezini ziyaret etmeye davet eden Demis Hassabis'e ulaştım. İki yıl önce MIT'de beni ziyaret ettiğinden beri ne kadar büyüdüklerine şaşırmıştım. Google onları 650 milyon dolara satın almıştı ve Demis'in “zekayı çözme” konusundaki cüretkar hedefinin peşinden giden parlak beyinlerle dolu geniş ofis ortamını görmek bana başarının gerçek bir olasılık olduğuna dair içgüdüsel bir his verdi.

Ertesi akşam arkadaşım Jaan Tallinn ile oluşturulmasına yardım ettiği yazılım Skype'ı kullanarak konuştum. FLI vizyonumuzu açıkladım ve bir saat sonra, bize bir şans vermeye karar verdi ve bizi yılda 100.000 $ 'a kadar finanse etti! Birisinin bana kazandığımdan daha fazla güven duyması kadar çok az şey bana dokunuyor, bu yüzden bir yıl sonra, Porto Riko konferansından sonra benim için dünya anlamına geliyordu.

1. bölümde bahsedildiği gibi, bunun şimdiye kadar yaptığı en iyi yatırım olduğu konusunda şaka yaptı.



Şekil 9.1: Ocak 2017 Asilomar konferansımız, Porto Riko devamı, AI ve ilgili alanlarda dikkate değer bir grup araştırmacıyı bir araya getirdi. Arka sıra, soldan sağa: Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin , Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán Ó hÉigeartaigh, Owain Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Firth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey,?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferriera, János Krámar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Rothberg. Ön sıra: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan Hadfield-Menell, Verity Harding, Shivon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo , ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt. Daha sonra geldi: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero. Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt . Daha sonra geldi: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero. Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt . Daha sonra geldi: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero. Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt . Daha sonra geldi: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero. Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Steph

Ertesi gün, yayıncım programımda bir boşluk bıraktı ve Londra Bilim Müzesi'ni ziyaretimle doldurdum. Uzun süre zekanın geçmişi ve geleceği hakkında takıntılı olduktan sonra, birdenbire düşüncelerimin fiziksel bir tezahürünün içinden geçtiğimi hissettim. Stephenon's Rocket'ten bilgi gelişimimizi temsil eden fantastik bir koleksiyon oluşturmuşlardı.

Model T Ford'a lokomotif, gerçek boyutlu bir Apollo 11 ay arazi aracı kopyası ve Babbage'nin "Fark Motoru" mekanik hesaplayıcısından günümüzün donanımına kadar uzanan bilgisayarlar. Ayrıca Galvano'nun kurbağa bacağı deneylerinden nöronlara, EEG'ye ve fMRI'ye kadar zihin anlayışımızın tarihi hakkında bir sergi vardı.

Nadiren ağlarım, ama çıkış yolunda ve South Kensington metro istasyonuna giderken yaya dolu bir tünelde yaptığım şey buydu. İşte tüm bu insanlar benim ne düşündüğümün farkında olmadan mutlu bir şekilde hayatlarını sürdürüyorlardı. İlk önce biz insanlar, bazı doğal süreçleri makinelerle nasıl kopyalayacağımızı, kendi rüzgarımızı ve şimşeklerimizi ve kendi mekanik beygir gücümüzü yaratmayı keşfettik. Yavaş yavaş vücudumuzun da makine olduğunu anlamaya başladık. Sonra sinir hücrelerinin keşfi, beden ve zihin arasındaki sınırı bulanıklaştırmaya başladı. Sonra sadece kaslarımızı değil, zihnimizi de aşan makineler üretmeye başladık. Öyleyse, ne olduğumuzu keşfetmeye paralel olarak, kendimizi kaçınılmaz olarak modası geçmiş hale mi getiriyoruz? Bu şiirsel olarak trajik olur.

Bu düşünce beni korkuttu ama aynı zamanda Yeni Yıl kararımı sürdürme kararlılığımı da güçlendirdi. FLI kurucularından oluşan ekibimizi tamamlamak için idealist genç gönüllülerden oluşan bir ekibe öncülük edecek bir kişiye daha ihtiyacımız olduğunu hissettim. Mantıksal seçim, Uluslararası Matematik Olimpiyatları'nda sadece gümüş madalya kazanmakla kalmayan, aynı zamanda daha büyük bir rol oynamak için sebep arayan bir düzine genç idealistin evi olan Citadel'i kuran parlak bir Harvard mezunu olan Viktoriya Krakovna idi. hayatları ve dünya. Meia ve ben onu, vizyonumuzu anlatmak için beş gün sonra evimize davet ettik ve suşiyi bitirmeden önce FLI doğmuştu.

Porto Riko Macerası

Bu, hala devam eden inanılmaz bir maceranın başlangıcı oldu. Bölüm 1'de bahsettiğim gibi, evimizde düzinelerce idealist öğrenci, profesör ve diğer yerel düşünürlerle düzenli beyin fırtınası toplantıları yaptık ve burada en yüksek puan alan fikirlerin projelere dönüştüğü - birinci bölümden Stephen'la yapay zekanın seçtiği ilk şey. Hawking, Stuart Russell ve Frank Wilczek, kamuoyundaki tartışmaları ateşlemeye yardımcı oldu. Yeni bir organizasyon kurmanın bebek adımlarına paralel olarak (şirket kurma, bir danışma kurulu işe alma ve bir web sitesi başlatma gibi), Alan Alda'nın teknolojinin geleceğini keşfettiği dolu bir MIT oditoryumunun önünde eğlenceli bir açılış etkinliği düzenledik. önde gelen uzmanlarla.



Şekil 9.2: Jaan Tallinn, Anthony Aguirre, gerçekten sizinki, Meia Chita-Tegmark ve Viktoriya Krakovna, 23 Mayıs 2014'te FLI'yi suşi ile birleştirmemizi kutluyor.

Yılın geri kalanında, birinci bölümde bahsettiğim gibi, dünyanın önde gelen yapay zeka araştırmacılarını yapay zekanın nasıl faydalı tutacağına ilişkin tartışmalara dahil etmeyi amaçlayan Porto Riko konferansını bir araya getirmeye odaklandık. Amacımız, yapay zeka güvenliği görüşmesini endişelenmekten çalışmaya kaydırmaktı: ne kadar endişeli olacağına dair tartışmalardan, iyi bir sonucun şansını en üst düzeye çıkarmak için hemen başlatılabilecek somut araştırma projeleri üzerinde anlaşmaya varmaya. Hazırlanmak için, dünyanın dört bir yanından gelecek vaat eden yapay zeka güvenliği araştırma fikirlerini topladık ve büyüyen proje listemiz hakkında topluluk geri bildirimi aradık. Stuart Russell ve bir grup çalışkan genç gönüllünün, özellikle Daniel Dewey, János Krámar ve Richard Mallah'ın yardımıyla, bu araştırma önceliklerini bir

konferansta tartışılacak belge. [1] Yapılacak çok sayıda değerli yapay zeka güvenliği araştırması olduğu konusunda fikir birliği oluşturmanın insanları bu tür araştırmalar yapmaya teşvik edeceğini umduk. Ayın nihai zaferi, birisini bunu finanse etmeye ikna edebilseydi bile olurdu, çünkü şimdiye kadar, hükümet finansman kurumlarından bu tür çalışmalar için esasen hiçbir destek yoktu.

Elon Musk'a girin. 2 Ağustos'ta, "Bostrom'dan" Okumaya değer Superintelligence "tweetiyle radarımıza çıktı. AI konusunda çok dikkatli olmamız gerekiyor. Nükleer bombalardan potansiyel olarak daha tehlikeli. " Ona ulaştım

Çabalarımız hakkında ve birkaç hafta sonra onunla telefonla konuşmalıyız. Kendimi oldukça gergin ve yıldızlardan etkilenmiş hissetmeme rağmen, sonuç olağanüstü idi: FLI bilimsel danışma kurulumuza katılmayı, konferansımıza katılmayı ve muhtemelen Porto Riko'da ilan edilecek ilk yapay zeka güvenliği araştırma programını finanse etmeyi kabul etti. Bu, FLI'de hepimizi heyecanlandırdı ve harika bir konferans oluşturma, gelecek vaat eden araştırma konularını belirleme ve onlar için topluluk desteği oluşturma çabalarımızı iki katına çıkardı.

İki ay sonra bir uzay sempozyumu için MIT'ye geldiğinde nihayet daha fazla planlama için Elon ile şahsen tanıştım. Binden fazla MIT öğrencisini bir rock yıldızı gibi büyülemesinden birkaç dakika sonra onunla küçük bir yeşil odada yalnız kalmak çok garip geldi, ancak birkaç dakika sonra tek düşünebildiğim ortak projemizdi. Onu anında sevdim. Samimiyet yayıyordu ve insanlığın uzun vadeli geleceğini ne kadar önemsediğinden ve özlemini cüretkar bir şekilde eyleme dönüştürdüğünden ilham aldım. İnsanlığın Evrenimizi keşfetmesini ve yerleşmesini istedi, bu yüzden bir uzay şirketi kurdu. Sürdürülebilir enerji istiyordu, bu yüzden bir güneş enerjisi şirketi ve bir elektrikli araba şirketi kurdu. Uzun boylu, yakışıklı, anlamlı ve inanılmaz derecede bilgili, insanların onu neden dinlediğini anlamak kolaydı.

Ne yazık ki, bu MIT olayı bana medyanın nasıl korku odaklı ve bölücü olabileceğini de öğretti. Elon'un sahne performansı, harika bir TV olacağını düşündüğüm uzay araştırmaları hakkında bir saatlik büyüleyici bir tartışmadan ibaretti. Sonunda, bir öğrenci ona yapay zeka hakkında konu dışı bir soru sordu. Cevabı, "yapay zeka ile iblisi çağırıyoruz" cümlesini içeriyordu. *sadece* çoğu medyanın bildirdiği şey - ve genellikle bağlam dışı. Birçok gazetecinin istemeden *tam zıttı* Porto Riko'da başarmaya çalıştığımız şey. Ortak zemini öne çıkararak toplumda fikir birliği oluşturmak isterken, medyanın bölünmeleri vurgulamak için bir teşviki vardı. Rapor edebilecekleri tartışma ne kadar fazlaysa, Nielsen puanları ve reklam gelirleri o kadar yüksek olur. Dahası, farklı görüş yelpazesinden insanların bir araya gelmelerine, birbirleriyle anlaşmalarına ve birbirlerini daha iyi anlamalarına yardımcı olmak isterken, medyada yer alan haberler farkında olmadan fikir yelpazesindeki insanları birbirlerine üzerek yanlış anlamaları yalnızca kulağa en kışkırtıcı olanlarını yayınlayarak körükledi. bağlamsız alıntılar. Bu nedenle, Porto Riko toplantısında gazetecileri yasaklamaya ve yasaklayan "Chatham Evi Kuralı" nı uygulamaya karar verdik.

katılımcılar daha sonra kimin ne dediğini açıklamaktan. *

Porto Riko konferansımız başarılı olmasına rağmen, kolay olmadı. Geri sayım çoğunlukla gayretli bir hazırlık çalışması gerektirdi, örneğin diğer katılımcıları çekmek için çok sayıda yapay zeka araştırmacısını arayarak veya atlayarak kritik bir katılımcı kitlesini bir araya getirdim ve aynı zamanda dramatik anlar da oldu - örneğin Aralık'ta sabah 7'ye kadar kalktığımda 27 Uruguay ile kötü bir telefon bağlantısıyla Elon'a ulaşmak için "Bunun işe yarayacağını sanmıyorum" söylendi. Bir yapay zeka güvenliği araştırma programının yanlış bir güvenlik duygusu sağlayabileceğinden endişeliydi, bu da pervasız araştırmacıların güvenliğe sözde hizmet verirken ilerlemelerine olanak tanıyordu. Ancak daha sonra, sesin durmadan kesilmesine rağmen, konuyu ana akım haline getirmenin ve daha fazla AI araştırmacısının AI güvenliği üzerinde çalışmasını sağlamanın büyük faydaları hakkında kapsamlı bir şekilde konuştuk. Çağrı kesildikten sonra, bana şimdiye kadarki en sevdiğim e-postalardan birini gönderdi: "Orada sonunda aramayı kaybettim. Her neyse, doktorlar iyi görünüyor. Araştırmayı üç yılda 5 milyon dolar ile desteklemekten mutluluk duyuyorum. Belki onu 10 milyon dolar yapmalıyız? "

Dört gün sonra, toplantıdan önce kısaca rahatlarken, havai fişeklerle aydınlatılmış bir Porto Riko sahilinde yeni yılda dans ederken, 2015 Meia ve benim için iyi bir başlangıç yaptı. Konferans da harika bir başlangıç yaptı: Daha fazla AI güvenliği araştırmasına ihtiyaç duyulduğu konusunda dikkate değer bir fikir birliği vardı ve konferans katılımcılarının daha fazla girdisine dayanarak, üzerinde çok çalıştığımız araştırma öncelikleri belgesinin iyileştirildi ve sonuçlandırıldı. 1. bölümdeki güvenlik araştırmasını onaylayan açık mektubu gözden geçirdik ve neredeyse herkesin imzalaması bizi çok mutlu etti.

Meia ve ben, hibe programımızın ayrıntılı planlarını kutsadığı otel odamızda Elon ile büyülü bir toplantı yaptık. Kişisel hayatı hakkında ne kadar gerçekçi ve samimi olduğu ve bize ne kadar ilgi duyduğu onu etkiledi. Bize nasıl tanıştığımızı sordu ve Meia'nın ayrıntılı hikayesini beğendi. Ertesi gün kendisiyle yapay zeka güvenliği ve neden

destekleyin ve her şey yolunda görünüyordu. 2

Konferansın zirvesi Elon'un bağış duyurusu 7 olarak planlandı
4 Ocak 2015 Pazar günü öğleden sonra, ve bu konuda o kadar gergindim ki bir gece önce uykuma dönüp döndüm. Ve sonra, gerçekleşeceği seansa gitmemizden sadece on beş dakika önce, bir engelle karşılaştık! Elon'un asistanı aradı ve Elon'un duyuruyu yapamayacak gibi göründüğünü söyledi ve Meia beni hiç bu kadar stresli ya da hayal kırıklığına uğramış görmediğini söyledi. Sonunda Elon geldi ve biz orada oturup konuşurken oturuma başlamak için geri sayım yapan saniyelerin sesini duyabiliyordum. Bunu açıkladı

Bir drone gemisine ilk etabın ilk başarılı inişini gerçekleştirmeyi umdukları çok önemli bir SpaceX roketinin fırlatılmasından sadece iki gün uzaktaydılar ve bu çok büyük bir dönüm noktası olduğundan, SpaceX ekibi bunu yapmak istemiyordu. onu içeren eşzamanlı medya sıçramalarıyla dikkatini dağıtın. Anthony Aguirre, her zamanki gibi havalı ve aklı başında, bunun şu anlama geldiğine işaret etti: *kimse* Bunun için medyanın ilgisini istedi, ne Elon ne de AI topluluğu. Moderatörlük yaptığım oturuma birkaç dakika geç geldik, ancak bir planımız vardı: duyurunun haber değeri taşımamasını sağlamak için hiçbir dolar tutarından bahsedilmeyecekti ve Elon'un duyurusunu gizli tutmak için Chatham House'u herkesin üzerine koyardım. inişin başarılı olup olmadığına bakılmaksızın, roketi uzay istasyonuna ulaştıysa dokuz gün boyunca dünyadan; roket fırlatıldığında patlarsa daha da fazla zamana ihtiyacı olacağını söyledi.

Duyuru için geri sayım nihayet sıfıra ulaştı. Moderatörlüğünü yaptığım süper zeka panelistleri, sandalyelerinde hâlâ yanımda oturuyorlardı: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg ve Vernor Vinge. İnsanlar yavaş yavaş alkışlamayı bıraktılar, ancak panelistler yerinde kaldı çünkü onlara nedenini açıklamadan kalmalarını söyledim. Meia daha sonra bana nabzının şu anda stratosfere ulaştığını ve Viktoriya Krakovna'nın sakinleştirici elini masanın altına sıkıştırdığını söyledi. Çalıştığımız, umduğumuz ve beklediğimiz anın bu olduğunu bilerek gülümsedim.

Toplantıda yapay zekanın yararlı olması için daha fazla araştırmaya ihtiyaç duyulduğuna dair böyle bir fikir birliği olduğu için çok mutluydum, dedim ve hemen üzerinde çalışabileceğimiz çok fazla somut araştırma yönü olduğunu söyledim. Ancak bu seansta ciddi risklerden söz edildiğini de ekledim, bu yüzden bara ve dışarıda düzenlenen konferans ziyafetine gitmeden önce moralimizi yükseltmek ve iyimser bir havaya girmek güzel olurdu. "Ve bu yüzden mikrofonu ... Elon Musk'a veriyorum!" Elon mikrofonu aldığında ve yapay zeka güvenlik araştırmalarına büyük miktarda para bağışlayacağını duyurduğunda tarihin yazılmakta olduğunu hissettim. Şaşırtıcı olmayan bir şekilde, evi yıktı. Planlandığı gibi, ne kadar olduğunu söylemedi, ama anlaştığımız gibi 10 milyon dolar olduğunu biliyordum.

Meia ve ben konferanstan sonra İsveç ve Romanya'daki ebeveynlerimizi ziyarete gittik ve nefes nefese, Stockholm'de babamla birlikte canlı yayınlanan roket fırlatmasını izledik. İniş girişimi maalesef Elon'un üstü kapalı bir şekilde RUD dediği şeyle sona erdi: "hızlı plansız sökme" ve

Başarılı bir okyanus inişini gerçekleştirmek, ekibinin on beş ayını daha aldı. [3]
Ancak, bizim gibi tüm uydular başarıyla yörüngeye fırlatıldı.

Elon tarafından bir tweet aracılığıyla milyonlarca takipçisine program veriyor. [4]

AI Güvenliğini Yaygınlaştırma

Porto Riko konferansının temel amacı, AI güvenliği araştırmasını yaygınlaştırmaktı ve bunun birden çok adımda ortaya çıktığını görmek heyecan vericiydi. İlk olarak, birçok araştırmacının, büyüyen bir akran topluluğunun parçası olduklarını fark ettiklerinde konuyla ilgilenirken kendilerini rahat hissetmeye başladıkları toplantının kendisi vardı. Birçok katılımcının cesaretlendirmesiyle derinden etkilendim. Örneğin, Cornell Üniversitesi AI profesörü Bart Selman bana e-posta göndererek, "Dürüst olmak gerekirse daha iyi organize edilmiş veya daha heyecan verici ve entelektüel olarak teşvik edici bilimsel toplantı görmedim" dedi.

Bir sonraki yaygınlaştırma adımı, 11 Ocak'ta Elon'un tweet atmasıyla başladı "Dünyanın en iyi yapay zeka geliştiricileri, yapay zeka için açık mektup imzaladı-
güvenlik araştırması " [5] kısa sürede sekiz binden fazla imzayı toplayan bir kayıt sayfasına bağlanarak, dünyanın en önde gelen yapay zeka oluşturucularının çoğu da dahil. Yapay zeka güvenliğinden endişe duyan insanların neden bahsettiklerini bilmediklerini iddia etmek birden bire zorlaştı, çünkü bu, lider YZ araştırmacılarından birinin kimin hakkında konuştuklarını bilmediğini ima etti. Açık mektup, dünyanın dört bir yanındaki medya tarafından, gazetecileri konferansımızdan men ettiğimiz için minnettar kılacak şekilde duyuruldu. Mektubun en korkutucu kelimesi "tuzaklar" olsa da, yine de cani sonlandırıcılar tarafından gösterilen "Elon Musk ve Stephen Hawking Robot Ayaklanmasını Önleme Umutlarında Açık Mektup İmzaladı" gibi manşetleri tetikledi. Gördüğümüz yüzlerce makaleden en sevdiğimizi diğerleriyle alay etmekti.

karmaşık, dönüştürücü teknoloji bir karnaval gösterisine dönüşüyor. " [6] Neyse ki, çok sayıda ayık haber makalesi de vardı ve bize başka bir zorluk da verdiler: Güvenilirliğimizi korumak için manuel olarak doğrulanması gereken ve "HAL 9000", "Terminatör" gibi şakaları ortadan kaldıran yeni imzaların seline ayak uydurmak, "" Sarah Jeanette Connor "ve" Skynet. " Bunun ve gelecekteki açık mektuplarımız için, Viktoriya Krakovna ve János Krámar, Jesse Galef, Eric Gastfriend ve Revathi Vinoth Kumar'ın vardiyalı olarak çalıştığı gönüllü bir dama tugayının örgütlenmesine yardımcı oldular, böylece Revathi Hindistan'da uyumaya gittiğinde, bayrağı teslim etti. Eric Boston'da vb.

Üçüncü anaakımlaştırma adımı dört gün sonra Elon'un bir bağlantıyı tweetlediği zaman başladı.

yapay zeka güvenliği araştırmasına 10 milyon dolar bağışta bulunduğunu duyurmamıza. [7] Bir hafta sonra, dünyanın her yerinden araştırmacıların başvurabileceği ve bu fon için rekabet edebileceği bir çevrimiçi portal başlattık. Başvuru sistemini bu kadar çabuk kırbaçlayabildik, çünkü Anthony ve ben geçen on yılı fizik hibeleri için benzer yarışmalar düzenleyerek geçirmiştik. Yüksek etkili bağışlara odaklanan Kaliforniya merkezli bir hayır kurumu olan Open Philanthropy Project, daha fazla hibe vermemize izin vermek için Elon'un hediyesini tamamlamayı cömertçe kabul etti. Konu yeni olduğu ve son teslim tarihi kısa olduğu için kaç başvuru alacağımızdan emin değildik. Yanıt, dünyanın dört bir yanından yaklaşık üç yüz ekibin yaklaşık 100 milyon dolar talep etmesiyle bizi uçurdu. Yapay zeka profesörlerinden ve diğer araştırmacılardan oluşan bir panel, teklifleri dikkatlice inceledi ve üç yıla kadar finanse edilen otuz yedi kazanan takımı seçti. Kazananların listesini açıkladığımızda, ilk kez medyanın faaliyetlerimize tepkisinin oldukça nüanslı olduğunu ve katil robot resimleri içermediğini gösterdi. Yapay zeka güvenliğinin boş bir konuşma olmadığı nihayet batıyordu: Yapılacak gerçekten faydalı işler vardı ve birçok harika araştırma ekibi bu çabaya katılmak için kolları sıvadı.

Dördüncü anaakımlaştırma adımı, önümüzdeki iki yıl içinde organik olarak gerçekleşti; çok sayıda teknik yayın ve dünya çapında AI güvenliği üzerine düzinelerce atölye çalışması, tipik olarak ana akım AI konferanslarının bir parçası olarak. Kalıcı insanlar uzun yıllar boyunca AI topluluğunu sınırlı bir başarı ile güvenlik araştırmalarına dahil etmeye çalıştılar, ancak şimdi işler gerçekten yükseldi. Bu yayınların birçoğu hibe programımız tarafından finanse edildi ve FLI olarak, bu atölye çalışmalarının olabildiğince çoğunu organize etmek ve finanse etmek için elimizden gelenin en iyisini yaptık, ancak bunların artan bir kısmı, kendi zamanlarına ve kaynaklarına yatırım yapan AI araştırmacıları tarafından sağlandı. Sonuç olarak, her zamankinden daha fazla sayıda araştırmacı, kendi meslektaşlarından güvenlik araştırmaları hakkında bilgi edindi ve yararlı olmanın yanı sıra, kafa karıştırmak için ilginç matematiksel ve hesaplama problemleri içeren eğlenceli de olabileceğini keşfetti.

Elbette karmaşık denklemler herkesin eğlence anlayışı değildir. Porto Riko konferansımızdan iki yıl sonra, Asilomar konferansımızdan önce FLI bursunu kazananlarımızın araştırmalarını sergileyebilecekleri bir teknik atölye çalışması gerçekleştirdik ve büyük ekranda matematiksel sembollerle slaytlar ardına slaytları izledik. Rice Üniversitesi'nde bir yapay zeka profesörü olan Moshe Vardi, toplantılar sıkıcı hale geldikten sonra bir yapay zeka güvenliği araştırma alanı kurmayı başardığımızı bildiği için şaka yaptı.

AI güvenliği çalışmasının bu çarpıcı büyümesi, akademi ile sınırlı değildi. Amazon, DeepMind, Facebook, Google, IBM ve Microsoft bir endüstri başlattı

faydalı AI için ortaklık. [8] Yapay zeka güvenliği bağışları, kâr amacı gütmeyen en büyük kardeş kuruluşlarımızda genişletilmiş araştırmalara olanak sağladı: Berkeley'deki Makine Zekası Araştırma Enstitüsü, Oxford'daki İnsanlığın Geleceği Enstitüsü ve Cambridge'deki (Birleşik Krallık) Varoluşsal Risk Araştırma Merkezi. 10 milyon ABD Doları veya daha fazla bağış, ek faydalı yapay zeka çabalarını başlattı: Cambridge'deki Leverhulme Zekanın Geleceği Merkezi, Pittsburgh'daki K&L Gates Etik ve Hesaplamalı Teknolojiler Vakfı ve Miami'deki Yapay Zeka Etik ve Yönetişim Fonu. Son olarak, bir milyar dolarlık taahhütle Elon Musk, San Francisco'da faydalı yapay zeka peşinde koşan kar amacı gütmeyen bir şirket olan OpenAI'yi başlatmak için diğer girişimcilerle ortaklık kurdu. Yapay zeka güvenliği araştırması kalıcı oldu.

Bu araştırma dalgalanmasıyla birlikte, hem bireysel hem de toplu olarak ifade edilen bir fikir dalgası geldi. AI üzerine endüstri Ortaklığı kuruluş ilkelerini yayınladı ve öneri listelerini içeren uzun raporlar ABD hükümeti, Stanford Üniversitesi ve IEEE (dünyanın en büyük teknik uzmanları kuruluşu) tarafından düzinelerce ile birlikte yayınlandı.

başka yerlerden alınan diğer raporlar ve pozisyon belgeleri. [9]

Asilomar katılımcıları arasında anlamlı bir tartışmayı kolaylaştırmak ve bu farklı topluluğun neyi kabul ettiğini öğrenmek istiyorduk. Lucas Perry bu nedenle bulduğumuz tüm belgeleri okumak ve onların fikirlerini çıkarmak gibi kahramanca bir görevi üstlendi. Anthony Aguirre tarafından başlatılan ve bir dizi uzun telecon tarafından sonuçlandırılan bir maraton çabasında, FLI ekibimiz daha sonra benzer görüşleri bir araya getirmeye ve gereksiz bürokratik lafları ortadan kaldırarak tek bir kısa ve öz ilkeler listesi çıkarmaya çalıştı. bu görüşmelerde ve başka yerlerde daha gayri resmi olarak ifade edilmiştir. Ancak bu liste hâlâ pek çok belirsizlik, çelişki ve yorumlama alanı içeriyordu, bu nedenle konferanstan önceki ay, katılımcılarla paylaştık, iyileştirilmiş veya yeni ilkeler için görüş ve önerilerini topladık. Bu topluluk girdisi, konferansta kullanılmak üzere önemli ölçüde revize edilmiş bir ilke listesi oluşturdu.

Asilomar'da liste iki aşamada daha da geliştirildi. İlk olarak, küçük gruplar en çok ilgilendikleri ilkeleri tartıştılar (şekil 9.4), ayrıntılı iyileştirmeler, geri bildirimler, yeni ilkeler ve eskilerin rakip versiyonlarını üretir. Son olarak, her bir ilkenin her versiyonu için destek düzeyini belirlemek için tüm katılımcıları araştırdık.



Şekil 9.3: Büyük beyin grupları Asilomar'daki AI ilkelerini düşünüyor.

Bu kolektif süreç hem kapsamlı hem de yorucuydu; Anthony, Meia ve ben, sonraki adımlar için gereken her şeyi zamanında derlemek için çabalarımızda konferansta uykumuzu ve öğle yemeğini kısıyoruz. Ama aynı zamanda heyecan vericiydi. Böylesine ayrıntılı, çetrefilli ve bazen çekişmeli tartışmalardan ve bu kadar geniş bir geri bildirim yelpazesinden sonra, bu son anket sırasında birçok ilkenin etrafında ortaya çıkan yüksek düzeydeki fikir birliğine hayret ettik ve bazıları% 97'den fazla destek aldı. Bu fikir birliği, nihai listeye dahil edilmek için yüksek bir çıta belirlememizi sağladı: yalnızca katılımcıların en az% 90'ının kabul ettiği ilkeleri tuttuk. Bu, bazı popüler ilkelerin

kişisel favorilerimden bazıları da dahil olmak üzere son dakikada düştü, [10] katılımcıların çoğunun, oditoryumun etrafından dolaştırdığımız kayıt formunda hepsini onaylayarak kendilerini rahat hissetmelerini sağladı. İşte sonuç.

Asilomar AI İlkeleri

Yapay zeka, şimdiden dünyanın her yerinden insanlar tarafından her gün kullanılan faydalı araçlar sağlamıştır. Aşağıdaki ilkelerin rehberliğinde devam eden gelişimi, önümüzdeki on yıllar ve yüzyıllarda insanlara yardım etmek ve onları güçlendirmek için inanılmaz fırsatlar sunacaktır.

R ARAŞTIRMA ben SSUES

§1 Araştırma Hedefi: *YZ araştırmasının amacı, yönsüz değil yaratmak olmalıdır zeka, ancak faydalı zeka.*

§2 Araştırma Fonu: *AI yatırımlarına finansman eşlik etmelidir bilgisayar bilimi, ekonomi, hukuk, etik ve sosyal bilimlerdeki çetrefilli sorular da dahil olmak üzere yararlı kullanımını sağlamaya yönelik araştırmalar için, örneğin:*

(a) *Gelecekteki yapay zeka sistemlerini ne kadar sağlam hale getirebiliriz ki, arızalanmadan veya hacklenmeden istiyoruz?*

(b) *Korurken otomasyon yoluyla refahımızı nasıl artırabiliriz? insanların kaynakları ve amacı?*

(c) *Hukuk sistemlerimizi daha adil ve verimli olacak şekilde nasıl güncelleyebiliriz? yapay zekaya ayak uydurmak ve yapay zeka ile ilişkili riskleri yönetmek?*

(d) *Yapay zeka hangi değerler dizisi ile uyumlu hale getirilmelidir ve hangi yasal ve etik statüsü olmalı mı?*

§3 Bilim-Politika Bağlantısı: *Yapıcı ve sağlıklı alışveriş olmalı AI araştırmacıları ve politika yapıcılar arasında.*

§4 Araştırma Kültürü: *İşbirliği, güven ve şeffaflık kültürü, YZ araştırmacıları ve geliştiricileri arasında teşvik edilmelidir.*

§5 Yarıştan Kaçınma: *AI sistemleri geliştiren ekipler, aktif olarak işbirliği yapmalıdır. güvenlik standartlarında köşe kesmekten kaçının.*

E THICS VE V ALUES

§6 Güvenlik: *AI sistemleri, operasyonları boyunca güvenli ve emniyetli olmalıdır ömür boyu ve uygulanabilir ve mümkün olduğu durumlarda doğrulanabilir.*

§7 Hata Şeffaflığı: *Bir AI sistemi zarar verirse, bunu yapmak mümkün olmalıdır nedenini araştırın.*

§8 Yargı Şeffaflığı: *Otonom bir sistemin herhangi bir katılımı Adli karar verme, yetkili bir insan otoritesi tarafından denetlenebilen tatmin edici bir açıklama sağlamalıdır.*

§9 Sorumluluk: *Gelişmiş yapay zeka sistemlerinin tasarımcıları ve kurucuları paydaşlar, kullanımlarının, yanlış kullanımlarının ve eylemlerinin ahlaki sonuçlarına, bu sonuçları şekillendirme sorumluluğu ve fırsatı ile birlikte.*

§10 Değer Hizalaması: *Son derece otonom yapay zeka sistemleri, amaç ve davranışlarının operasyonları boyunca insani değerlerle uyumlu olacağından emin olunabilir.*

§11 İnsani Değerler: *AI sistemleri, uygun şekilde tasarlanmalı ve çalıştırılmalıdır. insan onuru, haklar, özgürlükler ve kültürel çeşitlilik idealleriyle uyumludur.*

§12 Kişisel Gizlilik: *Kişilerin erişim, yönetme ve AI sistemlerinin bu verileri analiz etme ve kullanma gücü verildiğinde ürettikleri verileri kontrol edin.*

§13 Özgürlük ve Gizlilik: *AI'nın kişisel verilere uygulanması insanların gerçek veya algılanan özgürlüğünü makul olmayan bir şekilde kısıtlamak.*

§14 Paylaşılan Fayda: *AI teknolojileri, birçok kişiye fayda sağlamalı ve güçlendirmelidir mümkün olduğunca insan.*

§15 Paylaşılan Refah: *AI tarafından yaratılan ekonomik refah paylaşılmalıdır genel olarak, tüm insanlığın yararına.*

§16 İnsan Kontrolü: *İnsanlar nasıl ve nasıl yetkilendirileceğini seçmeli İnsan tarafından seçilen hedeflere ulaşmak için AI sistemlerine kararlar.*

§17 Yıkmama: *Son derece gelişmiş yapay zekanın kontrolünün sağladığı güç sistemler, toplum sağlığının dayandığı sosyal ve sivil süreçleri altüst etmek yerine saygı duymalı ve iyileştirmelidir.*

§18 AI Silahlanma Yarışı: *Ölümcül otonom silahlarda bir silahlanma yarışı, kaçınıldı.*

§19 Yetenek Dikkat: *Fikir birliği yok, güçlü olmaktan kaçınmalıyız gelecekteki AI yetenekleriyle ilgili üst sınırlarla ilgili varsayımlar.*

§20 Önemi: *Gelişmiş AI, tarihte büyük bir değişikliği temsil edebilir ve uygun bakım ve kaynaklarla planlanmalı ve yönetilmelidir.*

§21 Riskler: *AI sistemlerinin oluşturduğu riskler, özellikle yıkıcı veya varoluşsal riskler, beklenen etkileri ile orantılı planlama ve azaltma çabalarına tabi olmalıdır.*

§22 Özyinelemeli Kişisel Gelişim: *Özyinelemeli olarak kendi kendine Hızla artan kalite veya miktara yol açabilecek şekilde iyileştirme veya kendi kendini çoğaltma, sıkı güvenlik ve kontrol önlemlerine tabi olmalıdır.*

§23 Ortak Mal: *Süper zeka yalnızca hizmette geliştirilmelidir yaygın olarak paylaşılan etik ideallerin ve tek bir devlet veya kuruluştan ziyade tüm insanlığın yararına.*

İlkeleri çevrimiçi yayınladıktan sonra imza listesi çarpıcı bir şekilde büyüdü ve şimdiye kadar binden fazla AI araştırmacısı ve diğer birçok üst düzey düşünürden oluşan harika bir liste içeriyor. Siz de imza sahibi olarak katılmak istiyorsanız, bunu buradan yapabilirsiniz: http://futureoflife.org/ai-principles .

Sadece ilkeler hakkındaki fikir birliği düzeyine değil, aynı zamanda ne kadar güçlü olduklarına da şaşırdık. Elbette, bazıları ilk bakışta "Barış, sevgi ve annelik değerlidir" kadar tartışmalı geliyor. Ancak birçoğunun olumsuzluklarını formüle etmenin en kolay şekilde görülebileceği gibi, gerçek dişleri vardır. Örneğin, "Süper zeka imkansızdır!" §19'u ihlal ediyor ve "Yapay zeka kaynaklı varoluşsal riski azaltmaya yönelik araştırma yapmak tam bir israf!" §21'i ihlal ediyor.

Gerçekten de, YouTube'daki uzun vadeli panel tartışmamızı izlerseniz, kendiniz de görebileceğiniz gibi, [11] Elon Musk, Stuart Russell, Ray Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman ve Jaan Tallinn'in hepsi süper zekanın muhtemelen geliştirileceği ve güvenlik araştırmasının önemli olduğu konusunda hemfikirdi.

Asilomar AI İlkelerinin, sonuçta mantıklı AI stratejileri ve politikalarına yol açacak daha ayrıntılı tartışmalar için bir başlangıç noktası olarak hizmet edeceğini umuyorum. Bu ruhla, FLI medya direktörümüz Ariel Conn, Tucker'la çalıştı

Davey ve diğer ekip üyeleri, önde gelen yapay zeka araştırmacılarıyla ilkeler ve bunları nasıl yorumladıkları hakkında röportaj yaparken, David Stanley ve uluslararası FLI gönüllüleri ekibi ilkeleri temel dünya dillerine çevirdi.

Dikkatli İyimserlik

Bu sonsözün açılışında itiraf ettiğim gibi, hayatın geleceği konusunda uzun zamandır sahip olduğumdan daha iyimser hissediyorum. Nedenini açıklamak için kişisel hikayemi paylaştım.

Geçtiğimiz birkaç yıldaki deneyimlerim, iki ayrı nedenden dolayı iyimserliğimi artırdı. İlk olarak, AI topluluğunun, genellikle diğer alanlardan düşünürlerle işbirliği içinde, ilerideki zorlukları yapıcı bir şekilde üstlenmek için dikkate değer bir şekilde bir araya geldiğine şahit oldum. Elon, Asilomar toplantısından sonra bana yapay zeka güvenliğinin sadece birkaç yıl içinde sınır sorunlarından ana akıma geçişini şaşırtıcı bulduğunu söyledi ve ben de kendime hayran kaldım. Ve şimdi sadece 3. bölümdeki kısa vadeli sorunlar değil, aynı zamanda Asilomar AI İlkelerinde olduğu gibi süper zeka ve varoluşsal risk bile önemli tartışma konuları haline geliyor. Bu ilkelerin iki yıl önce Porto Riko'da benimsenmiş olmasının bir yolu yok, burada onu açık mektuba dönüştüren en korkutucu kulağa "tuzaklar" geliyordu.

İnsanları izlemeyi seviyorum ve Asilomar konferansının son sabahında bir noktada oditoryumun yanında durdum ve katılımcıların yapay zeka ve hukuk hakkında bir tartışmayı dinlemesini izledim. Şaşırtıcı bir şekilde, içimde sıcak ve bulanık bir his geçti ve birdenbire çok duygulandım. Bu Porto Riko'dan çok farklı geldi! O zamanlar, AI topluluğunun çoğunu saygı ve korku kombinasyonuyla gördüğümü hatırlıyorum - tam olarak karşıt bir ekip olarak değil, AI ile ilgili meslektaşlarım ve benim ikna etmemiz gerektiğini hissettiğimiz bir grup olarak. Ama şimdi o kadar bariz geldi ki hepimiz *aynı* takım. Muhtemelen bu kitabı okuyarak öğrendiğiniz gibi, yapay zeka ile nasıl harika bir gelecek yaratacağımıza dair cevapları hâlâ bilmiyorum, bu yüzden cevapları birlikte arayan büyüyen bir topluluğun parçası olmak harika bir duygu.



Şekil 9.4: Büyüyen bir topluluk, Asilomar'da birlikte yanıtlar aramaktadır.

İyimserleşmemin ikinci nedeni, FLI deneyiminin güçlendiriyor olması. Londra gözyaşlarımı tetikleyen şey bir kaçınılmazlık duygusuydu: rahatsız edici bir gelecek gelebilir ve bu konuda yapabileceğimiz hiçbir şey yoktu. Ama sonraki üç yıl kaderim karamsarlığımı çözdü. Bir grup parasız gönüllü bile zamanımızın tartışmasız en önemli sohbeti için olumlu bir fark yaratabilirse, o zaman birlikte çalışırsak ne yapabileceğimizi hayal edin!

Erik Brynjolfsson, Asilomar konuşmasında iki tür iyimserlikten bahsetti. İlk olarak, Güneş'in yarın sabah doğacağına dair olumlu beklenti gibi koşulsuz tür var. Sonra onun "dikkatli iyimserlik" dediği şey var, bu da dikkatlice planlar ve onlar için çok çalışırsanız iyi şeylerin olacağı beklentisidir. Hayatın geleceği hakkında şimdi hissettiğim türden bir iyimserlik bu.

Peki ne olabilir *sen* YZ çağına girerken yaşamın geleceği için olumlu bir fark yaratmak için ne yapmalı? Yakında açıklayacağım nedenlerle, eğer halihazırda değilseniz, dikkatli bir iyimser olmak için harika bir ilk adımın olduğunu düşünüyorum. Başarılı bir düşünceli iyimser olmak için, gelecek için olumlu vizyonlar geliştirmek çok önemlidir. MIT öğrencileri kariyer tavsiyesi için ofisime geldiklerinde, genellikle

On yıl içinde kendilerini nerede gördüklerini soruyorlar. Bir öğrenci "Belki de bir otobüs çarptıktan sonra bir kanser koğuşunda veya mezarlıkta olacağım" diye cevap verse, ona zor anlar yaşatırdım. Yalnızca olumsuz gelecekleri hayal etmek, kariyer planlamasına korkunç bir yaklaşımdır! Birinin çabalarının% 100'ünü hastalıklardan ve kazalardan kaçınmaya adaması, mutluluk için değil, hipokondri ve paranoya için harika bir reçetedir. Bunun yerine, hedeflerini coşkuyla tanımladığını duymak isterim, ardından tuzaklardan kaçınırken oraya ulaşmak için stratejileri tartışabiliriz.

Erik, oyun teorisine göre, evlilikler ve şirket birleşmelerinden bağımsız devletlerin ABD'yi kurma kararına kadar dünyadaki tüm işbirliğinin büyük bir kısmının temelini olumlu vizyonların oluşturduğuna dikkat çekti. Sonuçta, bunun sağlayacağı daha büyük kazancı hayal edemiyorsanız neden sahip olduğunuz bir şeyi feda edesiniz? Bu, sadece kendimiz için değil, aynı zamanda toplum ve insanlığın kendisi için de olumlu gelecekler hayal etmemiz gerektiği anlamına gelir. Başka bir deyişle, daha fazla varoluşsal ümide ihtiyacımız var! Yine de Meia'nın bana hatırlatmayı sevdiği gibi, Frankenstein'dan Terminatör'e, edebiyat ve filmdeki fütüristik vizyonlar ağırlıklı olarak distopiktir. Başka bir deyişle, biz bir toplum olarak geleceğimizi o varsayımsal MIT öğrencisi kadar zayıf planlıyoruz. Bu yüzden daha dikkatli iyimserlere ihtiyacımız var. *istemek* sadece ne tür bir geleceğin değil *korku,* planlamak ve üzerinde çalışmak için ortak hedefler bulabilmemiz için.

Bu kitap boyunca yapay zekanın bize nasıl büyük fırsatlar ve zorlu zorluklar sağladığını gördük. Esasen tüm AI zorluklarına yardımcı olması muhtemel bir strateji, birlikte hareket etmemizi ve insan toplumumuzu geliştirmemiz içindir. *önce* AI tamamen kalkıyor. Teknolojiye büyük güç aktarmadan önce, teknolojiyi sağlam ve faydalı hale getirmek için gençlerimizi eğitmemiz daha iyi. Teknoloji onları geçersiz kılmadan önce yasalarımızı modernize etmemiz daha iyi. Otonom silahlarda bir silahlanma yarışına dönüşmeden önce uluslararası çatışmaları çözmemiz daha iyi. Yapay zeka eşitsizlikleri potansiyel olarak artırmadan önce herkes için refah sağlayan bir ekonomi yaratmaktan daha iyiyiz. Yapay zeka güvenliği araştırma sonuçlarının göz ardı edilmek yerine uygulandığı bir toplumda daha iyiyiz. Ve ileriye baktığımızda, insanüstü YGZ ile ilgili zorluklara baktığımızda, bu standartları güçlü makinelere öğretmeye başlamadan önce en azından bazı temel etik standartları kabul etmemiz daha iyi. Kutuplaşmış ve kaotik bir dünyada, Yapay zekayı kötü niyetli amaçlar için kullanma gücüne sahip insanlar bunu yapmak için daha fazla motivasyona ve yeteneğe sahip olacak ve AGI oluşturmak için yarışan ekipler, iş birliği yapmaktansa güvenlik konusunda köşeleri kısmak için daha fazla baskı hissedecek. Özetle, paylaşıma yönelik işbirliği ile karakterize edilen daha uyumlu bir insan toplumu yaratabilirsek

hedefler, bu, AI devriminin iyi bitmesi olasılığını artıracaktır.

Diğer bir deyişle, yaşamın geleceğini iyileştirmenin en iyi yollarından biri yarını iyileştirmektir. Bunu birçok yönden yapma gücüne sahipsiniz. Elbette sandıkta oy kullanabilir ve politikacılarınıza eğitim, mahremiyet, ölümcül otonom silahlar, teknolojik işsizlik ve diğer konular hakkında ne düşündüğünüzü söyleyebilirsiniz. Ama aynı zamanda her gün ne satın almayı seçtiğiniz, hangi haberleri tüketmeyi seçtiğiniz, neyi paylaşmayı seçtiğiniz ve ne tür bir rol model olmayı seçtiğiniz üzerinden oy veriyorsunuz. Akıllı telefonunu kontrol ederek tüm konuşmalarını bölen biri mi, yoksa teknolojiyi planlı ve kasıtlı bir şekilde kullanarak güçlenmiş hisseden biri mi olmak istiyorsunuz? Teknolojinize sahip olmak mı yoksa teknolojinizin size sahip olmasını mı istiyorsunuz? Yapay zeka çağında insan olmanın ne anlama gelmesini istiyorsunuz? Lütfen tüm bunları çevrenizdekilerle tartışın

- bu sadece önemli bir sohbet değil, aynı zamanda büyüleyici bir sohbet.

Yapay zeka çağını şekillendirirken artık yaşamın geleceğinin koruyucularıyız. Londra'da ağlasam da, şimdi bu gelecek için kaçınılmaz bir şey olmadığını hissediyorum ve fark yaratmanın düşündüğümden çok daha kolay olduğunu biliyorum. Geleceğimiz taşa yazılmamış ve sadece başımıza gelmeyi beklemiyor - yaratmak bizim. Birlikte ilham verici bir tane yaratalım!

---

* Bu deneyim aynı zamanda haberleri kişisel olarak nasıl yorumlamam gerektiğini yeniden düşünmemi sağladı. Çoğu kuruluşun kendi siyasi gündemi olduğunun açıkça farkında olmama rağmen, artık siyasi olmayanlar da dahil olmak üzere tüm konularda merkezden uzak bir önyargıları olduğunu fark ettim.

Notlar

Bölüm 1

1. "AI Devrimi: Ölümsüzlüğümüz mü, Yok Olmamız mı?" *Bekle Ama Neden (* 27 Ocak 2015), http://waitbutwhy.com/2015/01/art artificial-intelligence-revolution-2.html .

2. Bu açık mektup, "Sağlam ve Yararlı Yapay Zeka için Araştırma Öncelikleri" bulunabilir. -de http://futureoflife.org/ai-open-letter/ .

3. Medyadaki klasik robot alarmı örneği: Ellie Zolfagharifard, "Yapay Zeka" Olabilir İnsanlığın Başına Gelen En Kötü Şey Olun '" *Günlük posta,* 2 Mayıs 2014; http://tinyurl.com/hawkingbots .

Bölüm 2

1. AGI teriminin kökeni hakkında notlar: http://wp.goertzel.org/who-coined-the-term-agi .

2. Hans Moravec, "Bilgisayar Donanımı İnsan Beynine Ne Zaman Uyacak?" *Journal of Evolution ve Teknoloji (* 1998), cilt. 1.

3. Yıla karşı hesaplama gücünü gösteren şekilde, 2011 öncesi veriler Ray Kurzweil'in kitabından alınmıştır. *Zihin Nasıl Oluşturulur,* ve sonraki veriler, içindeki referanslardan hesaplanır. https://en.wikipedia.org/wiki/FLOPS .

4. Kuantum hesaplamanın öncüsü David Deutsch, kuantum hesaplamayı nasıl gördüğünü anlatıyor: paralel evrenlerin kanıtı *Gerçekliğin Dokusu: Paralel Evrenlerin Bilimi - ve Etkileri (* Londra: Allen Lane, 1997). Kuantum paralel evrenleri dört çoklu evren seviyesinin üçüncüsü olarak ele almamı istiyorsanız, bunu önceki kitabımda bulacaksınız: Max Tegmark, *Matematiksel Evrenimiz: Gerçekliğin Nihai Doğası Arayışım (* New York: Knopf, 2014).

Bölüm 3

1. YouTube'da "Google DeepMind'in Derin Q-Öğrenme Oyun Atari Breakout" u şu adresten izleyin: https://tinyurl.com/atariai .
2. Bkz. Volodymyr Mnih ve diğerleri, "Derin Güçlendirmeli Öğrenme Yoluyla İnsan Seviyesinde Kontrol" *Doğa* 518 (26 Şubat 2015): 529–533, çevrimiçi olarak şu adresten ulaşılabilir: http://tinyurl.com/ataripaper .
3. İşte Big Dog robotunun çalıştığı bir video: https://www.youtube.com/watch?v=W1czBcnX1Ww .
4. AlphaGO'nun sansasyonel yaratıcılık çizgisi 5'e tepkiler için, bkz. “37. Lee Sedol vs AlphaGo Match 2 ", https://www.youtube.com/watch?v=JNrXgpSEEIE .
5. Demis Hassabis, insan Go oyuncularının AlphaGo'ya tepkilerini şöyle anlatıyor: https://www.youtube.com/watch?v=otJKzpNWZT4 .
6. Makine çevirisindeki son gelişmeler için bkz. Gideon Lewis-Kraus, "The Great AI Uyanış, " *New York Times Dergisi (* 14 Aralık 2016), çevrimiçi olarak şu adresten ulaşılabilir: http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html . GoogleTranslate şurada mevcuttur: https://translate.google.com .
7. Winograd Schema Challenge yarışması: http://tinyurl.com/winogradchallenge .
8. Ariane 5 patlama videosu: https://www.youtube.com/watch?v=qnHn8W1Em6E .
9. Ariane 5 Flight 501 Arıza raporu, soruşturma kurulu tarafından: http://tinyurl.com/arianeflop .
10. NASA'nın Mars Climate Orbiter Mishap Investigation Board Phase I raporu: http://tinyurl.com/marsflop .
11. Mariner 1 Venus görev başarısızlığına neyin sebep olduğuna dair en ayrıntılı ve tutarlı açıklama şuydu: tek bir matematiksel sembolün yanlış elle transkripsiyonu (eksik bir üst çubuk): http://tinyurl.com/marinerflop .
12. Sovyet Phobos 1 Mars görevinin başarısızlığının ayrıntılı bir açıklaması Wesley T'de bulunabilir. Huntress Jr. ve Mikhail Ya. Marov, *Güneş Sistemindeki Sovyet Robotları (* New York: Praxis Yayınları, 2011), s. 308.
13. Doğrulanmamış yazılımın 45 dakikada Knight Capital'e 440 milyon dolara mal olması: http://tinyurl.com/knightflop1 ve http://tinyurl.com/knightflop2 .
14. Wall Street "ani çöküşü" hakkında ABD hükümeti raporu: "6 Mayıs 2010 Pazar Olaylarına İlişkin Bulgular" (30 Eylül 2010), http://tinyurl.com/flashcra .
15. Binaların 3 boyutlu baskısı (https://www.youtube.com/watch?v=SObzNdyRTBs), mikromekanik cihazlar (http://tinyurl.com/tinyprinter) ve aradaki birçok şey (https://www.youtube.com/watch? v = xVU4FLrsPXs).
16. Topluluk tabanlı üretim laboratuvarlarının küresel haritası: https://www.fablabs.io/labs/map .
17. Robert Williams'ın endüstriyel bir robot tarafından öldürülmesiyle ilgili haber makalesi: http://tinyurl.com/williamsaccident .
18. Kenji Urada'nın endüstriyel bir robot tarafından öldürülmesiyle ilgili haber makalesi: http://tinyurl.com/uradaaccident .
19. Endüstriyel bir robot tarafından öldürülen Volkswagen işçisiyle ilgili haber makalesi: http://tinyurl.com/baunatalaccident .

20. İşçi ölümleriyle ilgili ABD hükümeti raporu: https://www.osha.gov/dep/fatcat/dep_fatcat.html .

21. Araba kazası ölüm istatistikleri: http://tinyurl.com/roadsafety2 ve http://tinyurl.com/roadsafety3 .

22. İlk Tesla otopilot ölümüyle ilgili olarak bkz. Andrew Buncombe, "Tesla Kazası: Açıkken Ölen Sürücü Otopilot Modu 'Harry Potter'ı İzliyordu' " *Bağımsız (* 1 Temmuz 2016), http://tinyurl.com/teslacrashstory . ABD Ulusal Karayolu Trafik Güvenliği İdaresi Kusurları Araştırma Ofisi raporu için bkz. http://tinyurl.com/teslacrashreport .

23. Daha fazlası için *Özgür Teşebbüsün Habercisi* felaket, bkz. RB Whittingham, *Suçlama Makinesi: İnsan Hatası Neden Kazalara Neden Oluyor (* Oxford, İngiltere: Elsevier, 2004).

24. Air France 447 kazasıyla ilgili belgesel: https://www.youtube.com/watch?v=dpPkp8OGQFI ; kaza raporu: http://tinyurl.com/af447report ; dış analiz: http://tinyurl.com/thomsonarticle .

25. 2003 ABD-Kanada karartmasına ilişkin resmi rapor: http://tinyurl.com/uscanadablackout .

26. Başkanın Three Mile Adasındaki Kaza Komisyonu'nun nihai raporu: http://www.threemileisland.org/downloads/188.pdf .

27. Yapay zekanın MRI tabanlı prostat kanseri teşhisinde insan radyologlarına nasıl rakip olabileceğini gösteren Hollanda çalışması: http://tinyurl.com/prostate-ai .

28. Yapay zekanın insan patologlarının akciğer kanseri teşhisinde en iyi nasıl olabileceğini gösteren Stanford çalışması: http://tinyurl.com/lungcancer-ai .

29. Therac-25 radyasyon tedavisi kazalarının araştırılması: http://tinyurl.com/theracfailure .

30. Panama'da kafa karıştırıcı kullanıcı arayüzünün neden olduğu ölümcül radyasyon doz aşımları hakkında rapor: http://tinyurl.com/cobalt60accident .

31. Robotik cerrahide istenmeyen olayların incelenmesi: https://arxiv.org/abs/1507.03518 .

32. Kötü hastane bakımı nedeniyle ölenlerin sayısına ilişkin makale: http://tinyurl.com/medaccidents .

33. Yahoo, kullanıcı hesaplarından bir milyarının (!) ihlal edildi: https://www.wired.com/2016/12/yahoo-hack-billion-users/ .

34. *New York Times* KKK katilinin beraatine ve sonradan mahkum edilmesine ilişkin makale: http://tinyurl.com/kkkacquittal .

35. Danziger ve diğerleri. 2011 çalışması (http://www.pnas.org/content/108/17/6889.full), aç yargıçların daha sert olduğunu savunan Keren Weinshall-Margela ve John Shapard tarafından kusurlu olduğu için eleştirildi (http://www.pnas.org/content/108/42/E833.full), ancak Danziger ve ark. iddialarının geçerli kalması konusunda ısrar ediyorlar (http://www.pnas.org/content/108/42/E834.full).

36. *Pro Publica* Suç işleme-tahmin yazılımında ırksal önyargı hakkında rapor: http://tinyurl.com/robojudge .

37. Denemelerde kanıt olarak fMRI ve diğer beyin tarama tekniklerinin kullanımı oldukça tartışmalıdır. bu tür tekniklerin güvenilirliği, birçok takım doğruluk oranlarının% 90'dan daha iyi olduğunu iddia etse de: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full .

38. PBS filmi yaptı *Dünyayı Kurtaran Adam* Vasili Arkhipov'un bekar olduğu olay hakkında Sovyet nükleer saldırısını elle engelledi: https://www.youtube.com/watch?v=4VPY2SgyG5w .

39. Stanislav Petrov'un bir ABD nükleer saldırısının uyarılarını yanlış alarm olarak nasıl reddettiğinin hikayesi filme dönüştü *Dünyayı Kurtaran Adam (* Bir önceki notta aynı adlı filmle karıştırılmamalıdır) ve Petrov, Birleşmiş Milletler'de onurlandırıldı ve Dünya Vatandaş Ödülü'ne layık görüldü: https://www.youtube.com/watch?v=IncSjwWQHMo .

40. Yapay zeka ve robotik araştırmacılarından otonom silahlar hakkında açık mektup: http://futureoflife.org/open-

mektup-özerk-silahlar / .

41. AU.S. görünüşte askeri bir AI silahlanma yarışı istiyor: http://tinyurl.com/workquote .

42. Amerika Birleşik Devletleri'nde 1913'ten beri servet eşitsizliğinin incelenmesi: http: // gabriel-zucman.eu/files/SaezZucman2015.pdf .

43. Oxfam küresel servet eşitsizliği raporu: http://tinyurl.com/oxfam2017 .

44. Teknoloji kaynaklı eşitsizlik hipotezine harika bir giriş için bkz.Erik Brynjolfsson ve Andrew McAfee, *İkinci Makine Çağı: Parlak Teknolojiler Zamanında İş, İlerleme ve Refah (* New York: Norton, 2014).

45. İçindeki makale *Atlantik Okyanusu* daha az eğitimli olanlar için düşen ücretler hakkında: http://tinyurl.com/wagedrop .

46. Çizilen veriler Facundo Alvaredo, Anthony B.Atkinson, Thomas Piketty, Emmanuel'ten alınmıştır. Saez ve Gabriel Zucman, *Dünya Servet ve Gelir Veritabanı (* http://www.wid.world), sermaye kazançları dahil.

47. James Manyika'nın, gelirin emekten sermayeye geçişini gösteren sunumu: http://futureoflife.org/data/PDF/james_manyika.pdf .

48. Oxford Üniversitesi'nden gelecekteki iş otomasyonu hakkında tahminler (http://tinyurl.com/automationoxford) ve McKinsey (http://tinyurl.com/automationm).

49. Robotik şefin videosu: https://www.youtube.com/watch?v=fE6i2OO6Y6s .

50. Marin Soljačić bu seçenekleri 2016 çalıştayı Computers Gone Wild: Impact and Yapay Zeka Alanındaki Gelişmelerin Toplum Üzerindeki Etkileri: http://futureoflife.org/2016/05/06/computers-gone-wild/ .

51. Andrew McAfee'nin nasıl daha iyi işler yaratılacağına dair önerileri: http://futureoflife.org/data/PDF/andrew_mcafee.pdf .

52. Teknolojik içerik için "bu sefer farklı" olduğunu savunan birçok akademik makaleye ek olarak işsizlik, "İnsanların Başvurması Gerekmiyor" videosu kısaca aynı noktaya işaret ediyor: https://www.youtube.com/watch?v=7Pq-S557XQU .

53. ABD Çalışma İstatistikleri Bürosu: http://www.bls.gov/cps/cpsaat11.htm .

54. Teknolojik işsizlik için "bu sefer farklı" argümanı: Federico Pistono, *Robotlar Olacak İşini Çal, ama Sorun Değil (* 2012), http://robotswillstealyourjob.com .

55. ABD at popülasyonundaki değişiklikler: http://tinyurl.com/horsedecline .

56. İşsizliğin refahı nasıl etkilediğini gösteren meta-analiz: Maike Luhmann ve diğerleri, "Öznel İyilik Hali ve Yaşamdaki Olaylara Uyum: AMeta-Analysis, " *Kişilik ve Sosyal Psikoloji Dergisi* 102, hayır. 3 (2012): 592; çevrimiçi olarak şu adresten temin edilebilir: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759 .

57. İnsanların refah duygusunu neyin artırdığına dair araştırmalar: Angela Duckworth, Tracy Steen ve Martin Seligman, "Klinik Uygulamada Pozitif Psikoloji" *Klinik Psikolojinin Yıllık Değerlendirmesi* 1 (2005): 629–651, çevrimiçi olarak http://tinyurl.com/wellb . Weiting Ng ve Ed Diener, "Zenginler ve Yoksullar İçin Ne Önemlidir? Dünya Çapında Öznel İyi Oluş, Mali Memnuniyet ve Postmaterialist İhtiyaçlar, " *Kişilik ve Sosyal Psikoloji Dergisi* 107, hayır. 2 (2014): 326, çevrimiçi

http://psycnet.apa.org/journals/psp/107/2/326 . Kirsten Weir, "İş Memnuniyetinden Daha Fazlası" *Psikoloji Üzerine İzleme* 44, hayır. 11 (Aralık 2013), çevrimiçi http://www.apa.org/monitor/2013/12/job- karşılamak.aspx .

58. Yaklaşık 10'u çarparak ^{11} nöronlar, yaklaşık 10^{4} nöron başına bağlantı ve yaklaşık bir (10^{0}) ateşleme

nöron başına her saniye yaklaşık 10 ^{15} FLOPS (1 petaFLOPS) bir insan beynini simüle etmek için yeterlidir, ancak ateşlemelerin ayrıntılı zamanlaması ve nöronların ve sinapsların küçük parçalarının da simüle edilmesinin gerekip gerekmediği sorusu dahil olmak üzere pek çok anlaşılmamış komplikasyon vardır. IBM bilgisayar bilimcisi Dharmendra Modha, 38 petaFLOPS gerektiğini tahmin etti (http://tinyurl.com/javln43), sinirbilimci Henry Markram birinin buna ihtiyaç duyduğunu tahmin ediyor

1.000 petaFLOPS (http://tinyurl.com/6rpohqv). AI araştırmacıları Katja Grace ve Paul Christiano, beyin simülasyonunun en maliyetli yönünün hesaplama değil, *iletişim,* ve bu da mevcut en iyi süper bilgisayarların yapabileceklerinin temelindeki bir görevdir:
http://aiimpacts.org/about .

59. İnsan beyninin hesaplama gücünün ilginç bir tahmini için: Hans Moravec "Ne zaman
Bilgisayar Donanımı İnsan Beynine Uygun mu? " *Journal of Evolution and Technology,* vol. 1 (1998).

4. Bölüm

1. İlk mekanik kuşun bir videosu için bkz. Markus Fischer, "ARobot That Flies Like a Bird," TED Konuşma, Temmuz 2011, https://www.ted.com/talks/a_robot_that_flies_like_a_bird .

Bölüm 5

1. Ray Kurzweil, *Tekillik Yakındır (* New York: Viking Press, 2005).
2. Ben Goertzel'in "Nanny AI" senaryosu burada açıklanmaktadır: https://wiki.lesswrong.com/wiki/Nanny_AI .
3. Makineler ve insanlar arasındaki ilişki ve makinelerin bizim olup olmadığı hakkında bir tartışma için köleler, bkz. Benjamin Wallace-Wells, "Boyhood," *New York* dergisi (20 Mayıs 2015), çevrimiçi olarak http://tinyurl.com/aislaves .
4. Akıl suçu Nick Bostrom'un kitabında tartışılıyor *Süper zeka* ve bu konuda daha teknik ayrıntıda son makale: Nick Bostrom, Allan Dafoe ve Carrick Flynn, "Policy Desiderata in the Development of Machine Superintelligence" (2016), http://www.nickbostrom.com/papers/aipolicy.pdf .
5. Matthew Schofield, "Stasi Renkli Almanların ABD Gözetleme Görüşüne Dair Anılar Programlar, " *McClatchy DC Bürosu (* 26 Haziran 2013), çevrimiçi olarak http://www.mcclatchydc.com/news/nation-world/national/article24750439.html .
6. İnsanların, kimsenin ulaşamayacağı sonuçlar yaratmaya nasıl teşvik edilebileceğine dair düşündürücü düşünceler için "Moloch Üzerine Meditasyonlar" ı öneriyorum http://slatestarcodex.com/2014/07/30/meditations-on- moloch .
7. Nükleer savaşın kaza sonucu başlamış olabileceği yakın görüşmelerin interaktif zaman çizelgesi için bkz. of Life Institute, "Kaza Sonucu Nükleer Savaş: Yakın Çağrıların Zaman Çizelgesi" çevrimiçi olarak http://tinyurl.com/nukeoops .
8. ABD nükleer testi mağdurlarına yapılan tazminat ödemeleri için bkz. ABD Adalet Bakanlığı web sitesi, "4/24/2015 Tarihine Kadar Ödüller" https://www.justice.gov/civil/awards-date-04242015 .
9. *Elektromanyetik Darbeden Amerika Birleşik Devletleri'ne Yönelik Tehdidi Değerlendirme Komisyonu Raporu (EMP) Saldırısı,* Nisan 2008, çevrimiçi olarak şu adresten ulaşılabilir: http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf .
10. Hem ABD hem de Sovyet bilim adamlarının bağımsız araştırmaları, Reagan ve Gorbaçov'u nükleer kış: PJ Crutzen ve JW Birks, "Bir Nükleer Savaş Sonrası Atmosfer: Öğlen Alacakaranlık" *Ambio* 11, hayır. 2/3 (1982): 114–125. RP Turco, OB Toon, TP Ackerman, JB Pollack ve C. Sagan, "Nükleer Kış: Çoklu Nükleer Patlamaların Küresel Sonuçları," *Bilim* 222 (1983): 1283–1292. VV Aleksandrov ve GL Stenchikov, "Nükleer Savaşın İklimsel Sonuçlarının Modellenmesi Üzerine" *Uygulamalı Matematik Üzerine Devam Etmek (* Moskova: SSCB Bilimler Akademisi Hesaplama Merkezi, 1983), 21. A. Robock, "Kar ve Buz Geri Bildirimleri Nükleer Kışın Etkilerini Uzatıyor" *Doğa* 310 (1984): 667–670.
11. Küresel nükleer savaşın iklim etkilerinin hesaplanması: A. Robock, L. Oman ve L. Stenchikov, "Nükleer Modern İklim Modeli ve Mevcut Nükleer Cephaneliklerle Yeniden Ziyaret Edildi: Hala Felaketli Sonuçlar, " *Jeofizik Araştırmalar Dergisi* 12 (2007): D13107.

Bölüm 6

1. Daha fazla bilgi için, bkz. Anders Sandberg, "Dyson Sphere SSS" http://www.aleph.se/nada/dysonFAQ.html .

2. Freeman Dyson'ın kendi adını taşıyan alanlarıyla ilgili çığır açan makalesi: Freeman Dyson, "Search for Artificial Kızılötesi Radyasyonun Yıldız Kaynakları, " *Bilim,* vol. 131 (1959): 1667–1668.

3. Louis Crane ve Shawn Westmoreland, önerilen kara delik motorunu "Are Black Hole Starships Possible ?, " http://arxiv.org/pdf/0908.1803.pdf .

4. CERN'den bilinen temel parçacıkları özetleyen güzel bir infografik için bkz. http://tinyurl.com/cernparticles .

5. Nükleer olmayan bir Orion prototipinin bu benzersiz videosu, nükleer bomba ile çalışan fikrini göstermektedir. roket itme gücü: https://www.youtube.com/watch?v=E3Lxx2VAYi8 .

6. Lazer yelkenciliğe pedagojik bir giriş: Robert L. Forward, "Roundtrip Interstellar Travel Lazer İtmeli Işık Raylarını Kullanma " *Uzay Aracı ve Roketler Dergisi* 21, hayır. 2 (Mart – Nisan 1984), çevrimiçi olarak şu adresten ulaşılabilir: http://www.lunarsail.com/Ligh .

7. Jay Olson, kozmik olarak genişleyen uygarlıkları "Homojen Kozmoloji ile Agresif Şekilde Genişleyen Medeniyetler, " *Klasik ve Kuantum Yerçekimi* 32 (2015), çevrimiçi olarak şu adresten ulaşılabilir: http://arxiv.org/abs/1411.4359 .

8. Uzak geleceğimizin ilk kapsamlı bilimsel analizi: Freeman J. Dyson, "Bitmeyen Zaman: Fizik ve Açık Bir Evrende Biyoloji, " *Modern Fizik İncelemeleri* 51, hayır. 3 (1979): 447, çevrimiçi olarak şu adresten ulaşılabilir: http://blog.regehr.org/extra_files/dyson.pdf .

9. Seth Lloyd'un yukarıda bahsedilen formülü, bize bir işlem sırasında bir hesaplama işlemi gerçekleştirmenin Zaman aralığı τ enerji $E \geq h / 4\tau$'ye mal olur, burada h Planck sabiti. Eğer istersek N Bir süre boyunca birbiri ardına (seri olarak) yapılan işlemler T, sonra $\tau = T/N$, yani $E/N \geq hN/4T$, bize performans sergileyebileceğimizi söyleyen $N \leq 2\sqrt{ET/h}$ enerji kullanan seri işlemler E ve zaman T. Yani hem enerji hem de zaman, bol miktarda olmasına yardımcı olan kaynaklardır. Enerjini ikiye bölersen n farklı paralel hesaplamalar, daha yavaş ve verimli çalışabilirler. $N \leq 2\sqrt{ETn/h}$. Nick Bostrom tahmin ediyor

   100 yıllık bir insan hayatını simüle etmenin yaklaşık $N = 10^{27}$ operasyonlar.

10. Yaşamın kökeninin neden çok nadir rastlanan bir tesadüf gerektirebileceğine dair dikkatli bir argüman görmek istiyorsanız,

    10 yaş üstü en yakın komşularımız 1000 birkaç metre ötede, Princeton fizikçisi ve astrobiyolog Edwin Turner'ın şu videosunu öneriyorum: "Olasılıksız Yaşam: Dünyadaki Yaşamın Kökeni İçin Çekici Olmayan Ama Makul Bir Senaryo" https://www.youtube.com/watch?v=Bt6n6Tu1beg .

11. Martin Rees'in dünya dışı zeka arayışı üzerine yazdığı makale: https://www.edge.org/annual-soru / 2016 / yanıt / 26665 .

Bölüm 7

1. Jeremy England'ın "dağıtıma dayalı adaptasyon" üzerine yaptığı popüler bir tartışma şu adreste bulunabilir: Natalie Wolchover, "ANew Physics Theory of Life", *Bilimsel amerikalı (* 28 Ocak 2014), çevrimiçi olarak şu adresten ulaşılabilir: https://www.scientifica . Ilya Prigogine ve Isabelle Stengers'in *Kaostan Çıkan Düzen: İnsanın Doğa ile Yeni Diyaloğu (* New York: Bantam, 1984) bunun temellerini atıyor.

2. Duygular ve fizyolojik kökenleri hakkında daha fazla bilgi için: William James, *Psikolojinin İlkeleri (* Yeni York: Henry Holt & Co., 1890); Robert Ornstein, *Bilincin Evrimi: Düşünme Şeklimizin Kökenleri (* New York: Simon & Schuster, 1992); António Damásio, *Descartes'Error: Duygu, Akıl ve İnsan Beyni (* New York: Penguin, 2005); ve António Damásio, *Kendilik Akla Geliyor: Bilinçli Beynin İnşası (* New York: Vintage, 2012).

3. Eliezer Yudkowsky, dost canlısı yapay zekanın hedeflerini mevcut hedeflerimizle değil, bizim *tutarlı tahmini irade (* CEV). Açıkça söylemek gerekirse, bu, daha fazlasını bilirsek, daha hızlı düşünürsek ve olmayı dilediğimiz insanlar olsaydık, idealize edilmiş bir versiyonumuzun ne isteyeceği olarak tanımlanır. Yudkowsky, CEV'yi 2004'te yayınladıktan kısa bir süre sonra eleştirmeye başladı (http://intelligence.org/files/CEV.pdf), hem uygulanması zor olduğu için hem de iyi tanımlanmış herhangi bir şeye yaklaşıp yaklaşmayacağı belirsiz olduğu için.

4. Ters pekiştirmeli öğrenme yaklaşımında, temel fikir, YZ'nin kendi kendi hedef-tatmini, ancak insan sahibinin. Bu nedenle, sahibinin ne istediği konusunda net olmadığında temkinli olmaya ve öğrenmek için elinden gelenin en iyisini yapmaya teşvik eder. Sahibinin onu kapatması da sorun değil, çünkü bu, sahibinin gerçekten ne istediğini yanlış anladığı anlamına gelir.

5. Steve Omohundro'nun yapay zeka hedefinin ortaya çıkmasıyla ilgili "Temel Yapay Zeka Sürücüleri" başlıklı makalesi çevrimiçi olarak şu adreste bulunabilir: http://tinyurl.com/omohundro2008 . Başlangıçta yayınlandı *Yapay Genel Zeka 2008: Birinci AGI Konferansı Bildirileri,* ed. Pei Wang, Ben Goertzel ve Stan Franklin (Amsterdam: IOS, 2008), 483–492.

6. Zeka körü körüne kullanıldığında ne olduğuna dair düşündürücü ve tartışmalı bir kitap etik temellerini sorgulamadan emirlere itaat edin: Hannah Arendt, *Kudüs'te Eichmann: Kötülüğün Sıradanlığı Üzerine Bir Rapor (* New York: Penguin, 1963). Eric Drexler'in yakın tarihli bir teklifi için de ilgili bir ikilem söz konusudur (http://www.fhi.ox.ac.uk/reports/2015-) süper zekayı, hiçbiri resmin tamamını anlamayan basit parçalara ayırarak kontrol altında tutmak. Eğer bu işe yararsa, bu yine, içsel bir ahlaki pusulaya sahip olmayan inanılmaz derecede güçlü bir araç sağlayabilir ve sahibinin her arzusunu herhangi bir ahlaki sıkıntı olmadan uygulayabilir. Bu, distopik bir diktatörlükteki bölümlere ayrılmış bir bürokrasiyi anımsatır: Bir kısım silahları nasıl kullanılacağını bilmeden üretir, bir kısmı mahkumları neden mahkum edildiklerini bilmeden infaz eder vb.

7. Altın Kural'ın modern bir varyantı, John Rawls'un varsayımsal bir durumun hiç kimse değilse adil olduğu fikridir. içinde hangi kişinin olacağını önceden bilmeden değiştirecekti.

8. Örneğin, Hitler'in üst düzey yetkililerinin birçoğunun IQ'ları oldukça yüksek bulundu. Nasıl olduğunu gör Nürnberg'de Yüksek Dereceli Üçüncü Reich Yetkililerinin IQ Puanları Doğru Muydu ?, " *Quora,* çevrimiçi olarak şu adresten temin edilebilir: http://tinyurl.com/nurembergiq .

Bölüm 8

1. Stuart Sutherland'ın bilinçle ilgili girişi oldukça eğlenceli: *Macmillan Sözlüğü Psikoloji (* Londra: Macmillan, 1989).

2. Kuantum mekaniğinin kurucu babalarından Erwin Schrödinger, kitabında bu açıklamayı yaptı. *Akıl ve Madde* düşünürken *geçmiş-* ve bilinçli yaşam ilk etapta asla gelişmemiş olsaydı ne olurdu. Öte yandan, yapay zekanın yükselişi, sahadaki boş sıralar için bir oyunla sonuçlanabileceğimiz mantıksal olasılığını artırıyor. *gelecek.*

3. *Stanford Felsefe Ansiklopedisi* farklı tanımların ve kullanımların kapsamlı bir incelemesini verir "bilinç" kelimesinin anlamı: http://tinyurl.com/stanfordconsciousness .

4. Yuval Noah Harari, *Homo Deus: Yarının Kısa Tarihi (* New York: HarperCollins, 2017): 116.

5. Bu, bir öncüden Sistem 1 ve Sistem 2'ye mükemmel bir giriş: Daniel Kahneman, *Hızlı ve Yavaş Düşünen (* New York: Farrar, Straus & Giroux, 2011).

6. Christof Koch'a bakın, *Bilinç Arayışı: Nörobiyolojik Bir Yaklaşım (* New York: WH Freeman, 2004).

7. Belki de beynimize giren bilginin çok küçük bir kısmının (örneğin 10-50 bit) farkındayızdır. her saniye: K. Küpfmüller, 1962, "Nachrichtenverarbeitung imMenschen," *Taschenbuch der Nachrichtenverarbeitung,* ed. K. Steinbuch (Berlin: Springer-Verlag, 1962): 1481–1502. T. Nørretranders, *Kullanıcı Yanılsaması: Bilincini Ölçüye Kadar Kesmek (* New York: Viking, 1991).

8. Michio Kaku, *Zihnin Geleceği: Anlama, Geliştirme ve Güçlendirme Bilimsel Arayışı akıl (* New York: Doubleday, 2014); Jeff Hawkins ve Sandra Blakeslee, *İstihbarat Üzerine (* New York: Times Books, 2007); Stanislas Dehaene, Michel Kerszberg ve Jean-Pierre Changeux, "Zahmetli Bilişsel Görevlerde Küresel Çalışma Alanının Nöronal Modeli" *Ulusal Bilimler Akademisi Bildiriler Kitabı* 95 (1998): 14529–14534.

9. Penfield'ın ünlü "Yanmış tostun kokusunu alabiliyorum" deneyini anlatan video: https://www.youtube.com/watch?v=mSN86kphL68 . Sensorimotor korteks ayrıntıları: Elaine Marieb ve Katja Hoehn, *Anatomi ve Fizyoloji,* 3. baskı (Upper Saddle River, NJ: Pearson, 2008), 391–395.

10. Bilincin sinirsel bağıntılarının (NCC'ler) incelenmesi, son yıllarda nörobilim topluluğu - bkz. örneğin, Geraint Rees, Gabriel Kreiman ve Christof Koch, "Neural Correlates of Consciousness in Humans," *Doğa Yorumları Nörobilim* 3 (2002): 261–270 ve Thomas Metzinger, *Bilincin Sinirsel İlişkileri: Ampirik ve Kavramsal Sorular (* Cambridge, MA: MIT Press, 2000).

11. Sürekli flaş bastırma nasıl çalışır: Christof Koch, *Bilinç Arayışı: A Nörobiyolojik Yaklaşım (* New York: WH Freeman, 2004); Christof Koch ve Naotsugu Tsuchiya, "Sürekli Flaş Bastırma Negatif Ardıl Görüntüleri Azaltır" *Doğa Sinirbilim* 8 (2005): 1096-1101.

12. Christof Koch, Marcello Massimini, Melanie Boly ve Giulio Tononi, "Sinirsel İlişkiler Bilinç: İlerleme ve Sorunlar " *Doğa Yorumları Nörobilim* 17 (2016): 307.

13. Koch'a bakın, *Bilinç Arayışı,* s. 260 ve daha fazla tartışma *Stanford Ansiklopedisi felsefe* http://tinyurl.com/consciousnessdelay .

14. Bilinçli algının senkronizasyonu üzerine: David Eagleman, *Beyin: Senin Hikayen (* Yeni York: Pantheon, 2015) ve *Stanford Felsefe Ansiklopedisi,* http://tinyurl.com/consciousnesssync .

15. Benjamin Libet, *Zihin Zamanı: Bilinçteki Zamansal Faktör (* Cambridge, MA: Harvard University Press, 2004); Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze ve John-Dylan Haynes, "İnsan Beyninde Özgür Kararların Bilinçsiz Belirleyicileri", *Doğa Sinirbilim* 11 (2008): 543–545, çevrimiçi olarak http://www.nature.com/neuro/journal/v11/n5/full/nn.2 .

16. Bilince yönelik son teorik yaklaşımlara örnekler:

- Daniel Dennett, *Bilinç Açıklandı (* Back Bay Books, 1992)
- Bernard Baars, *Bilinç Tiyatrosunda: Aklın Çalışma Alanı (* New York: Oxford University Press, 2001)
- Christof Koch, *Bilinç Arayışı: Nörobiyolojik Bir Yaklaşım (* New York: WH Freeman, 2004)
- Gerald Edelman ve Giulio Tononi, *Bir Bilinç Evreni: Madde Nasıl Hayal Gücüne Dönüşür (* New York: Hachette, 2008)
- António Damásio, *Kendilik Akla Geliyor: Bilinçli Beynin İnşası (* New York: Klasik, 2012)
- Stanislas Dehaene, *Bilinç ve Beyin: Beynin Düşüncelerimizi Nasıl Kodladığını Deşifre Etmek* (New York: Viking, 2014)
- Stanislas Dehaene, Michel Kerszberg ve Jean-Pierre Changeux, "Zahmetli Bilişsel Görevlerde Küresel Çalışma Alanının Nöronal Modeli" *Ulusal Bilimler Akademisi Bildiriler Kitabı* 95 (1998): 14529–14534
- Stanislas Dehaene, Lucie Charles, Jean-Rémi King ve Sébastien Marti, "Bilinçli İşlemenin Hesaplamalı Bir Teorisine Doğru" *Nörobiyolojide Güncel Görüş* 25 (2014): 760–784

17. David tarafından fizik ve felsefede "ortaya çıkış" teriminin farklı kullanımlarının kapsamlı tartışması Chalmers: http://cse3521.artifice.cc/Chalmers-Emergence.pdf .

18. Bilincin, belirli bir kompleks içinde işlenirken bilginin hissetme şekli olduğunu savunuyorum. yollar: https://arxiv.org/abs/physics/0510188 , https://arxiv.org/abs/0704.0646 Max Tegmark, *Matematiksel Evrenimiz (* New York: Knopf, 2014). David Chalmers, 1996 tarihli kitabında bununla ilgili bir düşünceyi ifade ediyor *Bilinçli Zihin:* "Deneyim içeriden gelen bilgidir; fizik, dışarıdan gelen bilgidir. "

19. Adenauer Casali ve ark., "A Heorik Temelli Bilinç İndeksi, Duyusaldan Bağımsız İşleme ve Davranış, " *Bilim Çeviri Tıbbı* 5 (2013): 198ra105, çevrimiçi http://tinyurl.com/zapzip .

20. Entegre bilgi teorisi sürekli sistemler için çalışmaz:

- https://arxiv.org/abs/1401.1219
- http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full
- https://arxiv.org/abs/1601.02626

21. Kısa süreli hafızası sadece yaklaşık 30 saniye olan Clive Wearing ile röportaj: https://www.youtube.com/watch?v=WmzU47i2xgw .

22. Scott Aaronson IIT eleştirisi: http://www.scottaaronson.com/blog/?p=1799 .

23. Cerrullo IIT eleştirisi, entegrasyonun bilinç için yeterli bir koşul olmadığını savunuyor: http://tinyurl.com/cerrullocritique .

24. Simüle edilen insanların zombi olacağına dair IIT tahmini: http://rstb.royalsocietypublishing.org/content/370/1668/20140167 .

25. Shanahan HTE eleştirisi: https://arxiv.org/pdf/1504.05696.pdf .

26. Kör görüş: http://tinyurl.com/blindsight-paper .

27. Belki de beynimize giren bilginin çok küçük bir kısmının (örneğin 10-50 bit) farkındayızdır. her saniye: Küpfmüller, “Nachrichtenverarbeitung imMenschen”; Nørretranders, *Kullanıcı Yanılsaması.*

28. "Erişimi olmayan bilinç" lehinde ve aleyhinde: Victor Lamme, "Sinirbilim Nasıl Olacak Bilinç Konusundaki Görüşümüzü Değiştirin, " *Bilişsel Sinirbilim (* 2010): 204–220, çevrimiçi olarak http://www.tandfonline.com/doi/abs/10.1080/17588921003731586 .

29. "Seçici Dikkat Testi" https://www.youtube.com/watch?v=vJG698U2Mvo .

30. Bkz. Lamme, "Nörobilim Bilinç Hakkındaki Görüşümüzü Nasıl Değiştirecek" n. 28.

31. Bu ve diğer ilgili konular Daniel Dennett'in kitabında ayrıntılı olarak tartışılmıştır. *Bilinç Açıkladı.*

32. Kahneman'a bakın, *Hızlı ve Yavaş Düşünen,* alıntı 5.

33. *Stanford Felsefe Ansiklopedisi* özgür irade tartışmasını inceler: https://plato.stanford.edu/entries/freewill .

34. Bir yapay zekanın neden özgür iradeye sahip gibi hissedeceğini açıklayan Seth Lloyd'un videosu: https://www.youtube.com/watch?v=Epj3DF8jDWk .

35. Steven Weinberg'e bakın, *Son Bir Teorinin Düşleri: Doğanın Temel Yasalarının Arayışı* (New York: Pantheon, 1992).

36. Uzak geleceğimizin ilk kapsamlı bilimsel analizi: Freeman J. Dyson, "Bitmeyen Zaman: Fizik ve Açık Bir Evrende Biyoloji, " *Modern Fizik İncelemeleri* 51, hayır. 3 (1979): 447, çevrimiçi olarak şu adresten ulaşılabilir: http://blog.regehr.org/extra_files/dyson.pdf .

Sonsöz

1. Açık mektup (http://futureoflife.org/ai-open-letter Porto Riko konferansından ortaya çıkan), AI sistemlerinin nasıl sağlam ve faydalı hale getirileceğine dair araştırmanın hem önemli hem de zamanında olduğunu ve bu araştırma öncelikleri belgesinde örneklendiği gibi bugün izlenebilecek somut araştırma yönlerinin olduğunu savundu: http://futureoflife.org/data/documents/research_priority.pdf .
2. Elon Musk ile yapay zeka güvenliğiyle ilgili video röportajım YouTube'da şu adreste bulunabilir: https://www.youtube.com/watch?v=rBw0eoZTY-g .
3. İşte hemen hemen tüm SpaceX roket iniş girişimlerinin güzel bir video derlemesi. ilk başarılı okyanus inişi: https://www.youtube.com/watch?v=AllaFzIPaG4 .
4. Elon Musk, yapay zeka güvenliği hibe yarışmamız hakkında tweetler: https://twitter.com/elonmusk/status/555743387056226304 .
5. Elon Musk, AI güvenliğini onaylayan açık mektubumuz hakkında tweetler: https://twitter.com/elonmusk/status/554320532133650432 .
6. Erik Sofge, "Yapay Zekadan Korkan Herkese Açık Mektup" (*Popüler Bilim,* 14 Ocak 2015), açık mektubumuzun korkutucu haber kapsamına alay ediyor: http://www.popsci.com/open-letter-everyone-tricked-fearing-ai .
7. Elon Musk, Future of Life Enstitüsü'ne ve yapay zeka güvenliği dünyasına yaptığı büyük bağış hakkında tweet attı araştırmacılar: https://twitter.com/elonmusk/status/555743387056226304 .
8. İnsanlara ve topluma fayda sağlamak için AI Ortaklığı hakkında daha fazla bilgi için web sitelerine bakın: https://www.partnershiponai.org .
9. Yapay zeka hakkında görüşleri ifade eden son raporlardan bazı örnekler: Yapay üzerine Yüz Yıllık Çalışma Zeka, 2015 Çalışma Paneli Raporu, "2030'da Yapay Zeka ve Yaşam" (Eylül 2016), http://tinyurl.com/stanfordai ; Yapay zekanın geleceği hakkında Beyaz Saray raporu: http://tinyurl.com/obamaAIreport ; Yapay Zeka ve işler hakkında Beyaz Saray raporu: http://tinyurl.com/AIjobsreport ; Yapay Zeka ve insan refahı üzerine IEEE raporu, "Etik Olarak Uyumlu Tasarım, Sürüm 1" (13 Aralık 2016), http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf ; US Robotics yol haritası: http://tinyurl.com/roboticsmap .
10. Son kestirmeyi yapmayan ilkeler arasında favorilerimden biri şuydu: "Bilinç Dikkat: Fikir birliği olmadığından, gelişmiş yapay zekanın bilinç veya duygulara sahip olup olmayacağı veya gerektirip gerektirmeyeceği konusunda güçlü varsayımlardan kaçınmalıyız. " Birçok yinelemeden geçti ve sonuncusunda, tartışmalı "bilinç" kelimesinin yerini "öznel deneyim" aldı - ancak yine de bu ilke yalnızca% 88 onay aldı ve% 90 sınırının çok az gerisinde kaldı.
11. Elon Musk ve diğer büyük beyinlerle süper zeka üzerine tartışma paneli: http://tinyurl.com/asilomarAI .

Penguin
Random
House